



INTRODUCING GRACE 2020 – A CLOUD COMPUTING BLUEPRINT



Bruce Yellin
Advisory Technology Consultant
EMC Corporation
Bruce.Yellin@EMC.Com

Table of Contents

Introduction	3
Mainframes	4
Timesharing.....	4
Transactional Computing.....	5
Grid Computing	5
In the Beginning.....	6
Three Forms of Cloud Computing – Private, Public, and Hybrid	7
The Crisis That Led Up to GRACE 2020	8
Introducing GRACE 2020.....	10
GRACE in More Depth – Security.....	12
GRACE in More Depth – Architecture	15
Application Virtualization	15
Transistor Technology.....	16
CPU Cores and Frequency	18
Memory and Cache Management	20
Storage	22
GRACE in More Depth – Securely Connecting with Any Device, Anywhere	25
GRACE in More Depth – The Labor Force	28
GRACE in More Depth – Form Factor	31
The End of Cell Towers?	33
GRACE Removes Barriers to 21st Century Cloud Computing.....	35
Vendor Lock-In, Pricing Models, and Service Level Agreements	36
Where were GRACEs Located?	37
GRACE’s Metadata Dashboard	38
Conclusion - Science Fiction Is a Prelude To Science Fact.....	41
Footnotes.....	44

Disclaimer: The views, processes, or methodologies published in this article are those of the author. They do not necessarily reflect EMC Corporation’s views, processes, or methodologies.

Introduction

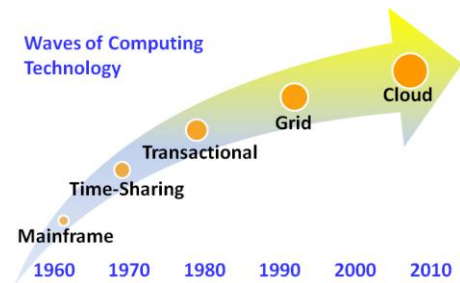
Sub-Commander T'Pol: The Vulcan Science Directorate has concluded that time travel is impossible.¹



Captain Jonathan Archer: Well, good for the Vulcan Science Directorate.

Star Trek: Enterprise's Captain Archer believes T'Pol is wrong and that time travel is possible. With “*future casting*”², as described by author Brian Johnson, time travel into the future can be achieved by combining science fiction with science fact. Science fiction writers blend their imagination with facts and make them believable stories. What if Captain Archer wanted to use science fiction to witness firsthand how cloud computing evolved? What would he see?

Traveling back to the mid-1960s, the Captain would have observed the birth of the fifth computing paradigm—cloud computing. With a lineage tracing back to earlier technologies, many say it didn't officially become a “wave” until 2008, while others claim it really gained traction in 2016. The fifth wave was based on earlier computing technology cycles—mainframes from the early 60's, timesharing in the 70's, transactional in the 80's, and grid computing in the 90's³.



As an agent of change often subject to resistance, cloud computing had a difficult birth. On one hand, this utility computing model offered users benefits such as high scalability, on-demand access, easy deployment, budget friendliness, and “greenness”. Its flaws came from integration issues, reliability, lack of standards, customization, security, management complexity, lock-in, performance, and others. CIOs wanted simplification and reduced computing costs. Consumers desired free and open computing, and were motivated, perhaps by marketing, to augment deskbound processing with intelligent mobile devices that leveraged feature-rich content.

What Captain Archer saw looking back in time was a computing paradigm that he took for granted in 2151, namely GRACE 2020. The GRACE platform would be recognized as the unification point for rival computing companies and the beginnings of true computer science harmony. New computing architectures were usually born from deficiencies and only become successful when technology became sufficiently advanced and affordable. During the latter part of the 20th century, expensive computing machines were severely underutilized. From 2011 through 2021, the number of physical and virtual servers grew 10X, data center information

grew 50X, and there would be a 75X increase in the number of files created⁴. Meanwhile, the worldwide number of IT professionals grew by less than 1.5X. Consumers of data processing services, using a myriad of devices, didn't care how or where the processing happened so long as it was fast and inexpensive. Paying homage to future casting, the year 2020 debuted a new approach to computing whose origins can be traced back to the early days of timesharing, yet were thoroughly infused with DNA-based security, virtual cloud computing philosophy, and even mainframe techniques. The Captain reviewed the four waves.

Mainframes

The 60's saw the emergence of the mainframe, and the most popular one in this wave was IBM's System/360. Among its claim to fame was that it was a family of systems, from a small S/360-20 to the large S/360-195, and that applications could run unchanged on any of them. These mainframes allowed users to submit a batch of FORTRAN or COBOL punched cards, created on an IBM 029 keypunch machine, into a card reader or remote-job entry (RJE) station miles away from the actual mainframe.



Clearly, this approach lacked efficiency. Millions of dollars of equipment would sit idle waiting for jobs to be loaded. A major breakthrough occurred in 1965 when the model 67 introduced “dynamic address translation” – a way to convert virtual memory addresses to physical memory addresses. It was a building block for the next technology wave.

Timesharing

Timesharing environments on the mainframe and minicomputer were part of the second wave. Through multi-tasking, multiprocessing, virtual memory, and other concepts, many users could access the same computer concurrently without interfering with each other, and made more efficient use of expensive machines. In the mid-70's, companies like National CSS offered timesharing services to thousands of users on a single virtual memory IBM mainframe. Using printing terminals, like the Teletype KSR33 or DECwriter LA 36, and dial-up analog modems, customers could access this “limitless” platform. “Punch card programmers” loved timesharing when they found they could achieve ten or more compiles a day using interactive terminals rather than one or two turnarounds a day with card decks. Their multi-tenant work was private and it seemed they had the entire mainframe to themselves.



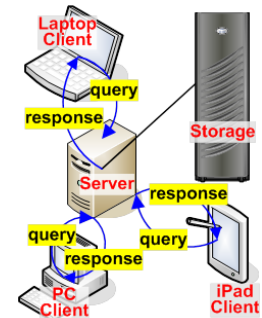
In the late 70's, National CSS sold a minicomputer for your private data center that could run the same programs that users ran on the timesharing mainframe, foreshadowing hybrid cloud computing. Marketed as a way to have predictable costs and load-balance work between local and remote timesharing like hybrid cloud computing, this NCSS 3200 machine ran the same mainframe NOMAD relational database system and could be accessed with the same terminals.



Transactional Computing

Transactional processing allows the requesting “client” program to send a single, indivisible transaction to a physically separated “server” program for processing. Representing the third wave of technology and forming the basis for most databases, a reliable network was needed to establish communications between the client and the server. An example of *online transaction processing* (OLTP) was the ATM machine – you entered information based on program prompts and the entire transaction was sent over a secure link to a server program requesting authorization for the client to dispense money. ATMs came into vogue in the early 80's.

To illustrate the difference between batch and transaction processing, think about a speech versus a conversation. Batch processing has thousands of transactions, similar to words in a speech, given without a response until the batch completes. Transaction processing is conversational, with a response given “immediately” to a query. This was also the basis for cloud computing.



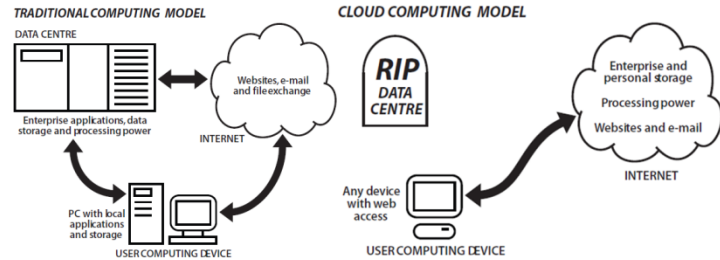
Grid Computing

Originating in the 90's, grid computing – the fourth wave - was a loosely coupled computing network that worked on pieces of the same program simultaneously. A middleware program distributed program parts to all of the participating machines. This approach was well suited to problems that could be parallelized and required a lot of processing. A popular example was SETI@Home's Search for Extraterrestrial Intelligence that analyzed radio telescope data.

Grid computing was exemplified by a distributed network of electric generators. Turning on a lamp, you don't know where the “utility” power comes from, nor if steam, coal, or wind powers the generators. Grid computing ran a program on geographically distributed computers of various sizes and operating systems. Cloud and grid computing were examples of “utility” computing – services that ran programs without regard to location, was available on-demand, was highly scalable, multitasked, and offered multi-tenancy.

In the Beginning...

Archer knew about the Internet, multi-tenancy, virtualization, self-service, and the grid, and took the world of private, public, and hybrid cloud computing for granted. Not defined simplistically by combining the features of a mainframe,



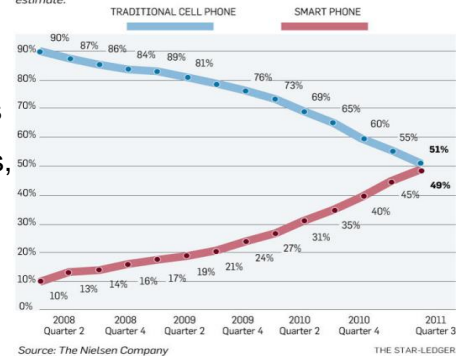
timesharing, transactional, or grid computing, its heritage was traced back to these earlier waves of technology and represented a major step towards utility computing. This illustration showed the shift from installing software and keeping data on your personal computer or enterprise computing platform, versus leveraging the world of cloud computing.⁵

Utility computing was highly efficient and paralleled the early days of electricity generation. Thomas Edison, inventor of the light bulb, also created “small radius” neighborhood direct current power stations. In contrast, Nikola Tesla, who worked for Edison in his New Jersey lab in 1884, favored George Westinghouse’s approach of alternating current since it could transmit electricity much further^{6 7}. Removing power stations from street corners and replacing them with rural, efficient generation plants revolutionized the delivery of electricity. When interconnected, a reliable and low cost electrical grid allowed people to depend on electricity and even take it for granted. “By 1907, utilities produced 40% of the power in the U.S. In 1920, that number stood at 70%, and a decade later, it was over 90%.⁸” Demand for electricity skyrocketed, and the same thing happened with cloud computing! There was no universal mandate or law that people use utility cloud computing, but the appeal was nonetheless indisputable.

The parallels between data processing and power station evolution were uncanny. In the 60’s, only top echelon organizations had mainframes. By the end of the decade and into the next, timesharing allowed more workers access to computers than ever thought possible. Personal desktops, miniaturized laptops, and eventually smart phones followed. In America, smart phones soon outsold traditional cell phones⁹. By 2013, cloud computing, like efficient power plants, made computing ubiquitous. Windows and Linux desktop PCs were passé and by 2020, they sat in museums next to VCRs, 8-track tapes, 35mm

TRADITIONAL VS. SMART PHONES

The number of smart phones owned by Americans is set to eclipse the number of traditional cell phones sometime in the next few months, according to a Nielsen estimate.



cameras, and rotating hard drives. Bandwidth was plentiful and thin-client computing re-emerged.

Some say it started when John McCarthy, a visionary like Tesla and Edison, said in 1961:

“If computers of the kind I have advocated become the computers of the future, then computing may someday be organized as a public utility just as the telephone system is a public utility....The computer utility could become the basis of a new and important industry.”¹⁰



Others believe J. C. R. Licklider, who as Director at ARPA wrote “Members and Affiliates of the Intergalactic Computer Network”¹¹, and described attributes of future cloud computing:

“If such a network as I envisage nebulously could be brought into operation, we would have at least four large computers, perhaps six or eight small computers, and a great assortment of disc files and magnetic tape units—not to mention the remote consoles and teletype stations—all churning away.”



And more recently, Dr. Eric Schmidt, the CEO of Google in 2006, discussed:

“...it starts with the premise that the data services and architecture should be on servers. We call it cloud computing – they should be in a ‘cloud’ somewhere. And that if you have the right kind of browser or the right kind of access, it doesn’t matter whether you have a PC or a Mac or a mobile phone or a BlackBerry or what have you – or new devices still to be developed – you can get access to the cloud...”¹²



So with a nod to future casting, let’s look at how the world of cloud computing in the year 2020 addressed some of the more vexing issues facing it in 2012. Back then, the computing world had three variations of cloud computing – private, public, and hybrid.

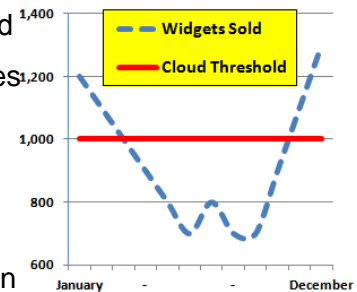
Three Forms of Cloud Computing – Private, Public, and Hybrid

The **private cloud** was simplistically defined as being dedicated to a single company or group – no one outside the organization could typically access these resources. The private cloud was part of a customized data center with dedicated management staffs and infrastructure, or privately hosted and managed in the cloud through third parties such as Amazon’s Private Cloud Service¹³ or Rackspace’s Managed Hosting Solutions¹⁴. Designed for elastic growth, these virtualized systems had dedicated resources and provided the same risk avoidance of traditional data centers since they were totally “walled off” from other organizations. Advocates of this technology, such as the New York Stock Exchange, employed private clouds so their “Customers can focus on developing proprietary advantage with their applications instead of worrying about the plumbing.”¹⁵

The **public cloud** including Google, Amazon, Yahoo!, Salesforce.com, and many others, was clearly the most popular model for consumers. The public used email, shared photos, purchased and stored music, backed up computers, and even did their taxes using the public cloud. These utility computing models were pay-as-you-go or free when subsidized by advertising, accessible worldwide, and most importantly, shared non-dedicated resources amongst tenants. They were highly automated, flexible, and built with uniform hardware and software, which helped keep OPEX and CAPEX costs down.

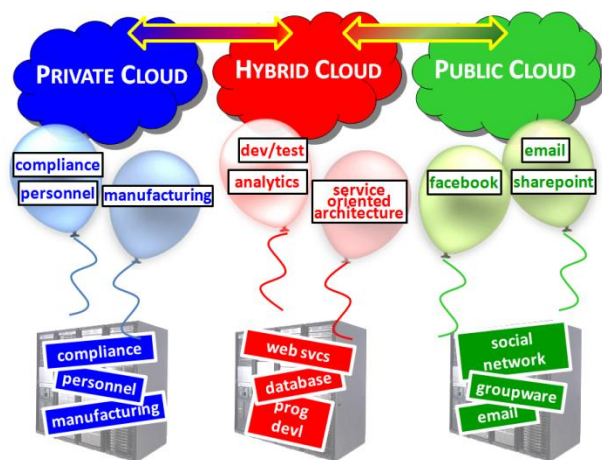
The logical blending of private and public clouds that allowed for the seamless flow of programs, data, and users from one cloud to the other based on demand was called the **hybrid cloud**.

This flexible model allowed for tight self-control for certain functions where it made business sense or for “split architectures” where perhaps data could be entered in the public cloud and processed in the private cloud. Other examples were found in retail organizations that had holiday data spikes. They opted to have just enough private cloud capacity for their “steady state” operations as illustrated by the **red line** in this graph and leveraged the public cloud during peak customer buying periods when the **blue dashed line** went above the red line.



Similar to the earlier NCSS 3200 minicomputer versus timesharing example, a business decision was made to off-load some of their transactions from November to February into the elastic public cloud.

The corporate landscape wrestled with the matchup of applications and clouds, and which should run where. There were arguments for ROI and privacy, data center power dilemmas, fiefdoms, and even dramatic changes in business policies and regulations. End-users had a much easier time of it – their decisions were based on cost or lack thereof, coolness, and quickness.

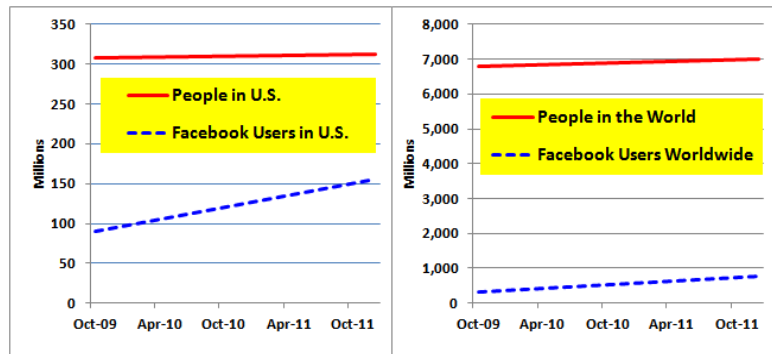


The Crisis That Led Up to GRACE 2020

It became evident that trying to keep data pristine and private was a full time effort, incredibly expensive, and possibly futile. Growing threefold every year, Google reported that 1.3% of the searches they processed were infected, and “...Google's anti-malware team uncovered more

than 3 million potentially harmful Web sites.”¹⁶. In 2010, “Symantec discovered 286 million new and unique threats from malicious software, or about nine per second, up from 240 million in 2009. The company said that the amount of harmful software in the world passed the amount of beneficial software in 2007...”¹⁷. Not only was the threat vividly present, but so was the threat rate. “95,000—that’s the number of malware pieces analyzed by SophosLabs every day in 2010, nearly doubling the number of malware pieces we tracked in 2009. This accounts for one unique file every 0.9 seconds...”¹⁸.

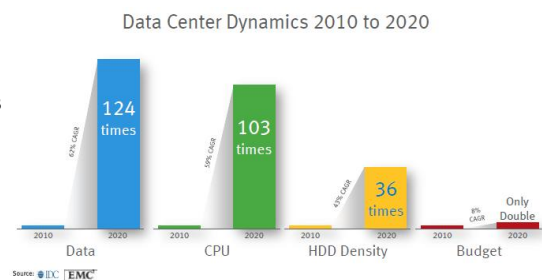
Despite the growing threats, users flocked to the cloud and the data explosion continued unabated. Wikipedia contained over 10 million articles in 2011 in 273 different languages¹⁹. The social networking giant Facebook supported 300 million users in 2009²⁰ and over ¾ of



a billion people just two years later²¹ - i.e., 30 times faster than the population growth! Skype served up more than a half a billion calling minutes per day with 41% being video calls, all the while handling 25% of all international calling minutes worldwide in 2011.²² Young people were moving away from email and towards Twitter, text messaging, and instant messages.

Unscrupulous criminals targeted this prey-rich environment since users had passwords like “123456”, “password”, and “qwerty”. In fact, “clickjacking” became a popular way of scamming an individual’s identity through false pages and postings. Data was harvested and more passwords were stolen through keylogging and screen scraping as users innocently went about their day online. To data thieves, sharing content was like feeding a fire with gasoline. Users demanded a safe place to interact with each other, yet no serious progress was made.

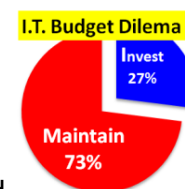
Businesses watched these events in horror and adamantly refused to put their crown jewels into the cloud. They feared tarnishing or destroying their corporate interests, and did not assume the risk just to save some money. Meanwhile, 2010 forecasts showed data and CPU growth would increase over 100 fold and hard drive density 36 fold²³ at a time when IT budget merely doubled. The attacks against local data centers continued and the business



cost of IT further drained precious resources in the midst of a decade-long global recession. Computer vendors were increasingly pushing cloud computing as the “next best thing”.

By 2013, IT organizations and end-users had reached a tipping point. Computer systems left and right were experiencing attacks from cybercriminals, computer viruses and worms, theft, breaches, spyware, WikiLeaks, botnets, spamming, malware, phishing, denial of service, data loss, disasters, running out of space, Internet “cookie” abuse, registry issues, incompatibilities, vendor wars, end-of-line products, upgrading, computer crashes, buggy software, rebooting, poor documentation, backups, fragmentation, and dozens of other glitches, annoyances, and disruptions. Data processing was unsustainable. Just like Steve Jobs knew when a product was just right and not merely good enough, users demanded more from the cloud. While they appreciated the efforts of the Cloud Standards Customer Council²⁴, Open Data Center Alliance²⁵, Cloud Security Alliance²⁶, and other groups, users around the world grew more impatient with the slow pace of change. As a tribute to America’s founding fathers, a large group of users gathered in Philadelphia and drafted a computing Bill of Rights.

Businesses were under great economic pressure to create competitive advantages through IT at the same time IT struggled to meet corporate demand with over 73% of its budget spent to maintain existing systems. IT was becoming a corporate “boat anchor”. Companies like McDonalds needed to sell more hamburgers and not spend as much money on data processing. The future looked bleak. The industry’s opinion was that unless the paradigm changed, data growth alone would swamp the data center.



Introducing GRACE 2020

Captain Archer saw the real power of GRACE’s architecture as the combination of private, public, and hybrid cloud concepts with a generous helping of “warp plasma” allowing it to:

- maintain total security and privacy
- offer unbounded dynamic scaling
- be virus-proof and immune to hacking
- offer rich reporting and decision making
- allow resources to fluidly move based on policies or on-demand
- be professionally maintained

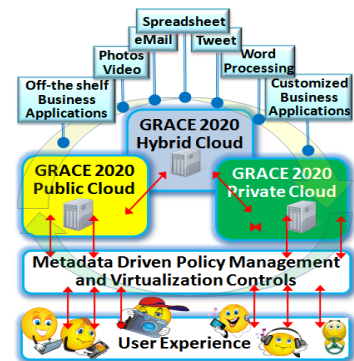
Just seven short years after Dr. Schmidt’s comments, the world witnessed the birth of a new computing paradigm whose heritage was deeply entrenched in the proven ways of the past,

together with the Bill of Rights that addressed all the computing wrongs of the last half-century. Thus was born the Geographical Area of Computing Excellence, or GRACE for short.

Named for U.S. Rear Admiral Grace Hopper (pictured, right), a computing pioneer who helped spearhead the development of COBOL, GRACE became the ultimate computing architecture and a true revolutionary step forward.

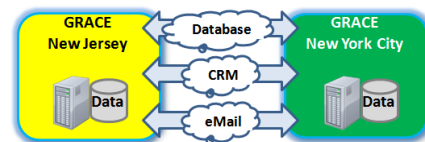


No longer science fiction but science fact, GRACE leveraged transparent virtualization, the enormous power of 3-D multi-core processors, and geographic storage over giant, wide area data pipelines. These systems benefitted from single-instancing, compression, and encryption, along with a healthy dose of earlier NUMA and GRID computing innovation that yielded a “mainframe” concept light years ahead of the 60’s version.



While GRACE leveraged hardware-assisted virtualization concepts from 2013, it took another seven years until GRACE 2020 was ready. The basic premise was utility computing. In other words, the consumer did not care where the computing power came from or how it was created – a concept the world had only dreamed about.

- GRACE was truly stateless and portable. Applications and data could easily flow globally between various forms of GRACE all with a single login.
- The GRACE network employed advanced wide area communication technology that automatically deduped, compressed, and encrypted data traffic. With most of the world using the cloud, the deduplication/compression rate was very high, often nearing 100:1.
- GRACE provided the highest level of disaster recovery and business continuance, especially during times of natural disasters when massive workloads could be shifted to safer geographic locations, either automatically or on demand.
- GRACE was elastic. Applications could use more or less resources by leveraging metadata profiles.
- GRACE supported all user devices and provided an uncompromised experience.
- GRACE was seamless to the user.
- GRACE was affordable (or free) to the user and saved companies money.



- GRACE built on the work of The Cloud Computing Interoperability Forum²⁷, the Open Cloud Manifesto²⁸, the Open Cloud Computing Interface²⁹. and others to structure a totally interoperable environment that fostered near instant transferability of users' data and programs, thereby avoiding the dreaded vendor lock-in.

Companies loved it because they could truly focus on their business and not have to spend valuable resources on data center technology. Hardware and software vendors built GRACE-certified products to defined specifications agreed to by the GRACE committee, similar to the “plug-compatible” mainframes and de facto standards of the 70's. The evils of everyday computing that became prevalent in 2012 disappeared.

GRACE in More Depth – Security

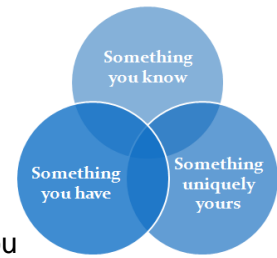
With one of every ten Internet downloads containing damaging code³⁰, one of GRACE's most significant achievements was the infusion of computing safety in our everyday lives. It also marked the beginning of the end of burdensome security precautions.

A 2013 committee was chartered to design a new security architecture to alleviate the world's fear of entrusting data to nebulous cloud entities. From a security aspect, the model had to be at a minimum as secure as data would be in the private cloud. Data scientists, having learned a lot about computer break-ins that led to theft, viruses, and other attacks, decided to blend the utmost in human security with the most trusted data security model. They devised an identity management system that fundamentally combined human DNA binary coding with the common two-factor authentication and created a widely accepted three-factor authentication.



Their guiding principles focused on air-tight identity identification, which required individuals to have proper credentials, certificates, and proof of who they said they were. Before a user could access any cloud service, they needed a painless and unobtrusive vetting by their police department or other government office using either the enhanced E-Verify system³¹ or through a passport verification check. This process, which assigned everyone a unique DNA sequencing number, allowed them to legitimately use the cloud based on assigned roles. While initially a bit draconian, people came to regard this like airport screening, and gradually supported the project. They were assured that their *number* would not be used for unethical purposes. Clearly, anyone unable to pass the initial security check would be relegated to the “no cloud compute privileges” list, similar to the “don't fly” list used at the beginning of the 21st century.

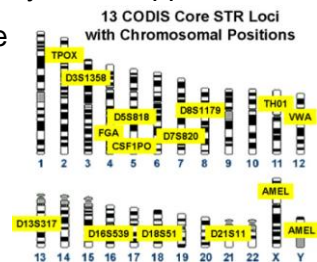
Authentication proves your identify to a system. With proof and a defined profile, a system allows you or your surrogate access to data or programs – e.g., only Finance staffers can access financial records. It is achieved through something you **know** like a password, something you **have** like an ATM card, and something uniquely **yours** like your fingerprint. The more factors present, the greater the likelihood that you are who you really say you are. Single factor authentication, often relying on a password, was risky since it could be the same password you used everywhere or simplistic like “123abc” or “password”. There was no guarantee you were who you said you were.



Two factor authentication used a password and often a numeric sequence that changed frequently. With two factors, the likelihood of identity theft was reduced, but not eliminated. Its weakness was its reliance on the password, but little could be done about it since it was hard to remember unique, random sequences. People often wrote them down or used words from a poem or song. Numeric values likely came from a 6-digit RSA token code that changed them every 60 seconds. A system then looked up the agreed password and combined it with the code at the second it was entered to determine who you really were with some certainty.



Adding a third unique factor greatly enhanced security, and that’s precisely what happened. In 2014, data scientists and biometric engineers developed an inexpensive and fast way to add the third factor - “Touch DNA”. Rather than taking blood samples or saliva from users, they leveraged the work done by federal customs agents who encode Touch DNA on everyone’s passport. Modeled after the algorithms employed by the FBI’s CODIS (Combined DNA Index System) database, security scientists leveraged the STR (Short Tandem Repeats) Loci (different data points)³² to create that third unique authentication factor.

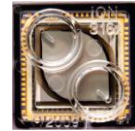


Similar to fingerprints, every creature on Earth has a unique DNA sequence. For identification purposes, you didn’t need to examine the entire 3 billion base pair genome, just 13 chromosomes repeated at multiple locations in a DNA string³³.



This let you know if two STR’s were from the same person. The design used existing touch DNA from fraud-proof international passports, touch DNA birth records, or through a simple local law enforcement registration process. The signature was kept in an encrypted database for a user’s cloud access and as a surrogate for unique data certificates that could be used for data security enforcement.

In 2010, Dr. Jonathan Rothberg of Ion Torrent introduced a chip called the “Personal Genome Machine”³⁴. This miniaturized breakthrough decoded genome sequences that previously needed complicated and expensive equipment to perform.



What was missing was a low-cost, high speed, miniature touch DNA sensor that could be built into a handheld device, and five years later, scientists leveraging Dr. Rothberg’s work developed it. Touch DNA could be logged by pressing the on-off button of a device or the touch-sensitive smart-phone screen. The device mathematically encoded and transmitted the digitally unique string to the computing service, along with any other two-factors for identification. This cleared one of the last remaining hurdles – a security mechanism was not a hindrance to legitimate cloud computing. It was used in conjunction with a single sign-on to allow ease of flow in the clouds.



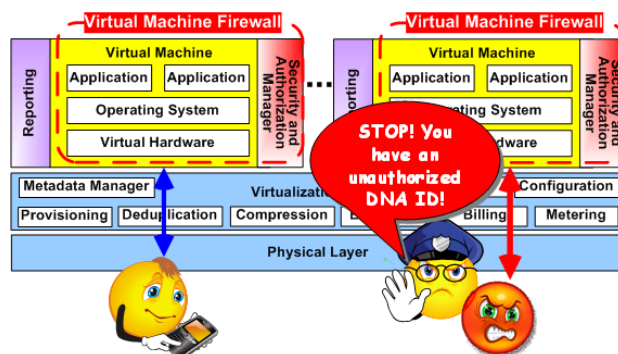
This approach created an iron-clad, safe environment where cloud and program access was granted only if you truly had permission. For example, should a known data criminal create a virus, their DNA ID must be part of it and it would be blocked. Their cloud access would have already been suspended. Legitimate programs were screened by GRACE authorities and assigned DNA IDs, allowing authorized programs to be used by authorized users.

Devices were assigned signatures, so if they were stolen, they were put on a “don’t compute list”. With GPS tracking in devices like iPads, unusual access attempts were tracked down and the suspect detained. Theft dropped dramatically since the stolen devices were rendered useless. The system was fool-proof and could even tell genetically identical twins apart.



By 2020, GRACE was “always and forever” accountable, so compliance issues became “black and white”. The court system could track data access and stop many frivolous lawsuits. Like an employee’s ID card, the DNA ID allowed access to GRACE, but once in, it only permitted access to authorized programs and data. The world quickly ran out of data criminals – no more hacking, identity theft, stolen data, computer viruses, denials of service, and so forth.

Security inside GRACE’s architecture was built at the virtual machine level. Machines were protected by a “Virtual Machine Firewall” leveraging information in the



Security and Authorization Manager to ensure iron-clad protection. The firewall was the traffic cop that allowed information to flow into and out of the virtual application and aided with compliance.

GRACE in More Depth – Architecture

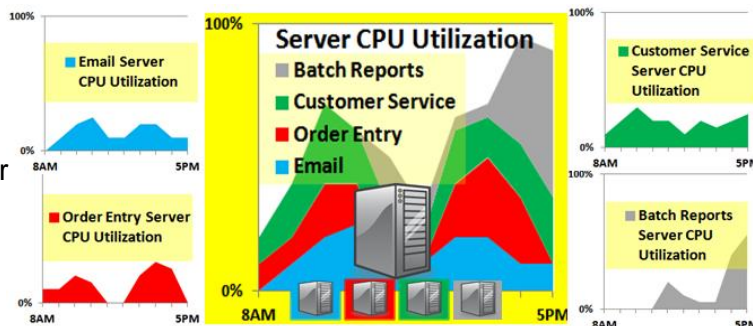
Let's explore the virtualization, processor (transistor, CPU cores and clock frequency, cache system), and disk subsystems used back then in a GRACE cloud frame.

Application Virtualization

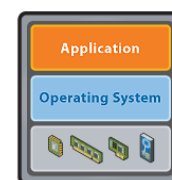
When you are called a “spitting image”, it means you look like someone else; perhaps one of your parents. GRACE was the spitting image of parts of previous computer models. At a high level, GRACE was based on virtualization made popular by VMware in the late 90's. Itself an extension to some early mainframes and timesharing designs, VMware abstracted and virtualized the x86 architecture to

achieve higher machine utilization.

This illustrates four groups with underutilized servers. Combining their workloads into a single virtualized server, they all co-exist and still achieve their performance goals.

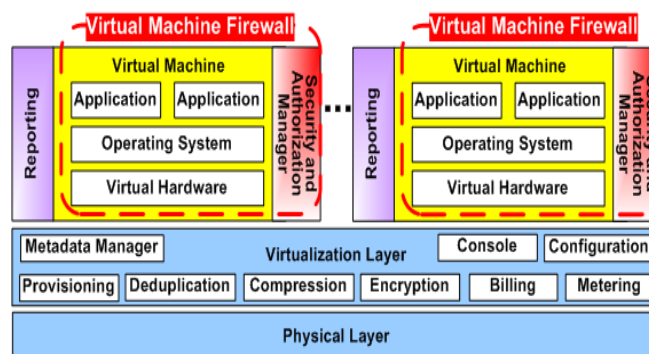


Virtualization separates physical hardware from the application/operating system, and allows them to use virtualized CPU, memory, and network interfaces as though it was on a similar physical machine.³⁵ You then extend this to run many applications on the same server by decoupling applications and operating environments from metal components, allowing organizations to exploit expensive, underutilized hardware so fewer physical servers were needed. It saved power, simplified support, achieved greater returns on investment, load balanced applications, and reacted faster to changing events.



A VMware virtual machine

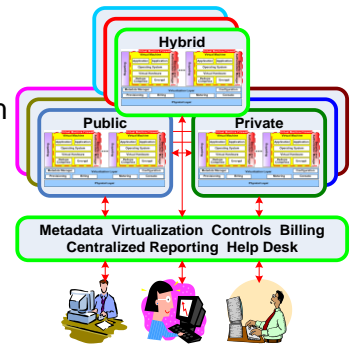
GRACE's “industrial strength” virtualization was integrated with data deduplication, compression, encryption, security and authorization, provisioning, reporting, metering, configuration, billing, console



operations, and universal operating system APIs. The metadata manager tracks policies and the security mechanism makes GRACE tamper-proof.

The 2018 GRACE committee created a standard design for public, private, and hybrid compute systems that proved fundamental to the cloud experience. Through advanced metadata enforcement, if one virtual program needed to “talk” with another program, the “true identity” of both programs was verified based on DNA ID to ensure integrity and a common, agreed set of APIs allowed bidirectional flow and interaction, even between the clouds.

Programs moved transparently between them based on their metadata permissions. When a process’s response time slowed and policies permitted, additional inter- and intra-cloud resources were brought on-line, allowing programs and data in different systems to effectively act as one, similar to active-active clustering behavior from 2012.



Transistor Technology

The miracle Ted Hoff invented in 1971 – the microprocessor – a CPU on a chip – would not have been possible without William Shockley, John Bardeen, and Walter Brattain’s invention of the transistor in 1947³⁶.



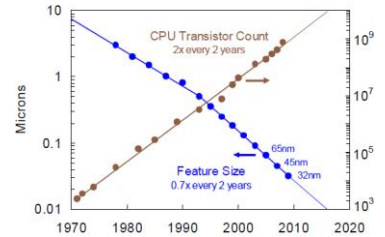
The Intel 4004, shown on the right by Mr. Hoff, was a 4-bit machine that contained 2,300 transistors. The next breakthrough was their 8008 with a 50% increase in transistors.

The transistor opens and closes by electrical charge and was a fundamental CPU building block. More transistors generally meant a faster processor, although bus design, architecture, clock speed, etc. were also factors. This chart shows the transistor counts rapidly increasing over time³⁷. “Processing power, measured in millions of instructions per second (MIPS), has steadily risen because of increased transistor density coupled with improved multiple core processor microarchitecture.”³⁸ The Intel 22nm Westmere-EX chip had over 2 billion transistors, with miniaturization approaching the atomic level -

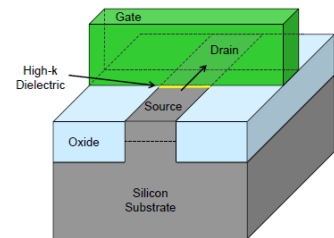
Year	Intel Processor	# Transistors
1971	4004	2,300
1972	8008	3,500
1974	8080	4,500
1978	8086	29,000
1982	80286	134,000
1985	80386	275,000
1989	80486	1,200,000
1993	Pentium	3,100,000
1995	Pentium Pro	5,500,000
1997	Pentium II	7,500,000
1999	Pentium III	9,500,000
2001	Pentium 4	42,000,000
2002	Pentium M	55,000,000
2006	Core 2 Duo	291,000,000
2007	Quad-Core Xeon	820,000,000
2009	Nehalem	731,000,000
2010	Westmere	1,170,000,000
2010	Nehalem-EX	2,300,000,000
2011	Westmere-EX	2,600,000,000

over 6 million transistors could fit into the period at the end of this sentence³⁹. Intel boldly predicted that technology from their Many Integrated Core (MIC) Architecture by 2020 could supply the market with ExaFLOP/s performance - “...hundreds times more than today’s fastest supercomputers.”⁴⁰

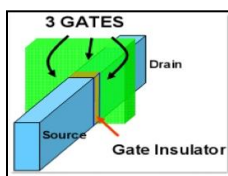
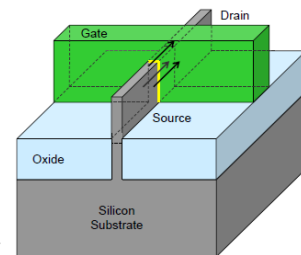
Gordon Moore's famous prediction demonstrates the near linear logarithmic growth of transistors through 2010⁴¹. Doubling the number of transistors every two years meant ever faster processors as chip components shrunk in size and approached atomic levels. Early chip designers felt the world would someday see a 10GHz processor, but they "...discovered that they would get so hot it would melt through the Earth..."⁴². Sadly, one of the drawbacks to increased miniaturization was increased heat output. Shrinking silicon transistors increased electrical resistance and "leakage" since the two opposing surfaces had less contact area. Chips that leak need more power which also means they generate more heat. Heat dissipation was a problem with densely packed components since there was less surface area to dissipate the higher heat output, which meant bigger heat sinks, larger fans, or the need for liquid cooling.



A major Intel breakthrough in 2011 allowed for the continuation of Moore's prediction. The "Tri-Gate" or "3-D" transistor was one of the most significant developments in its 60+ year history.⁴³ Flat transistors had three critical areas – source, gate, and drain. "When you apply a positive voltage to the gate, it attracts the few free electrons in the positively-charged substrate to the gate's oxide layer. This creates an electron channel between the source and drain terminals. If you then apply a positive voltage to the drain, electrons will flow from the source through the electron channel to the drain."⁴⁴ The transistor is "ON" when the gate is charged; otherwise it is "OFF".



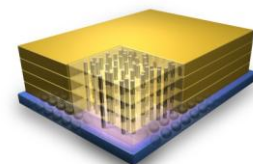
The tri-gate transistor had three dimensions for the conducting channel in the silicon part of the transistor instead of just one on top⁴⁵. This dramatically increases the surface area of the conducting channel⁴⁶.



Larger surfaces translate into needing less power compared to the flat transistor. The greater surface contact also reduces leakage, which means a far lower heat profile. In the end, the Tri-Gate operates almost 3X faster and runs cooler

with less power than its predecessor.

That same year, IBM and Micron introduced the 3-D "Hybrid Memory Cube"⁴⁷. It provided 128 GB/s performance while regular 2-D memory ran at roughly 1/10 that rate, due in part to the advantages of shorter

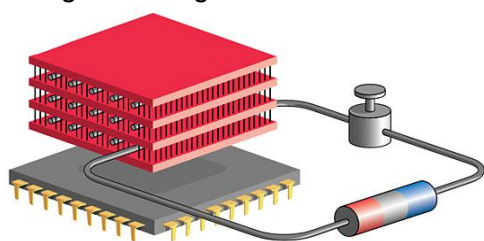


electrical paths. It used 70% of the power in a footprint 90% smaller than traditional memory designs.

Circuitry continued to shrink, aided by a breakthrough by UCLA Berkeley researchers. Their 2011 exploration of ferroelectric materials resulted in a way to lower the voltage needed to store a charge in a capacitor⁴⁸, which reduced power requirements and produced less wasteful heat. Circuits reached 14nm in 2013, 10nm two years later, and were 7nm in size by 2017.⁴⁹

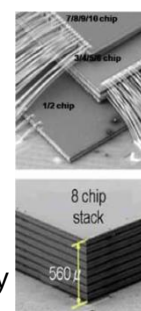
Year	Intel Model	# Transistors
2013	Chip2013	5,200,000,000
2015	Chip2015	10,400,000,000
2017	Chip2017	20,800,000,000
2019	Chip2019	41,600,000,000
2021	Chip2021	83,200,000,000

Leveraging the popular tri-gate transistor and memory cube, the next innovation was the three dimensional processor. Based on early 3-D attempts shown to the right⁵⁰, the 3-D design on the left had advanced sufficiently to allow a multi-level, multi-core chip to emerge. Cooling the device was addressed short term with micro-miniature tubes of



circuitry refrigerant⁵¹. The longer term electrohydrodynamic approach to cooling these tiny, hot circuits was invented in 2011 by Jeff Didion, a NASA thermal engineer, and Jamal Seyed-Yagoobi, a professor at the

Illinois Institute of Technology, who found a way to cool electronics in outer space.⁵² The electrohydrodynamic design is still used on the Enterprise.



Similar to the economies of scale constructing a 10-story, 50 apartment high-rise building versus 50 individual “ranch” houses, 3-D processors with shorter circuit paths became lightning fast and easy to manufacture. Moore’s Law continued to hold true and even accelerate a bit as GRACE used chips with over 50 million transistors in 2020.

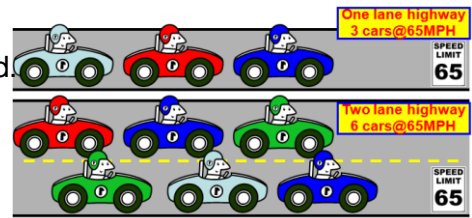
CPU Cores and Frequency

Processors not only had more transistors in 2020, they also had a greater number of processor cores per chip. Through mid-2005, everyone used single core processors running at ever increasing CPU frequencies. As with dense transistor packaging, increasing frequencies increased heat, which became an issue for the CPU industry and threatened to stop Moore’s law in its tracks. If these single core processors went any faster, the server itself could overheat.

In the same general size of a single core processor, IBM's 2001 POWER4 dual-core processor breakthrough advanced Moore's Law. Previously, a server had two discrete processors in what was called Symmetric Multiprocessing (SMP). SMP could now be achieved in a single chip!

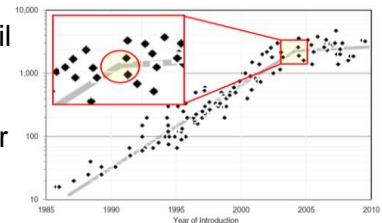
In May, 2005, AMD introduced its 233 million transistor dual core processor, and in July that year, Intel announced their version with 291 million transistors. While each core operated at a somewhat reduced CPU frequency compared to the latest single core designs, it could handle twice the workload. It was like expanding a single lane highway into a two lane highway while holding the speed limit to 65 MPH.– the highway handled

twice as many cars as long as they all went the same speed. Rather than run a single core perhaps at 3GHz, you could operate two cores at 2GHz in the same CPU socket for nearly 4GHz of equivalent power.

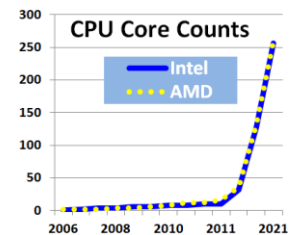


For the next five years, Intel “raced” AMD for bragging rights on how many compute cores they could fit into a socket. Intel quad-core processors were available in October, 2006 and AMD followed suit eleven months later. Six cores were sold by Intel in September, 2008 and by AMD in June, 2009. In March, 2010, both companies shipped eight cores, and AMD debuted twelve cores at that time. Intel announced ten cores in April, 2011. AMD shipped its 16 core “Bulldog” processor in late 2011⁵³. By 2014, both companies had 32 core offerings. What a race!

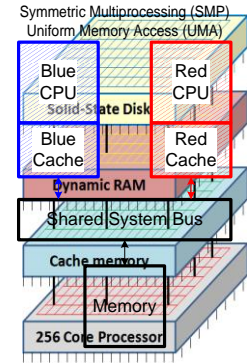
CPU frequency kept to Moore's pace, doubling every two years until late 2004⁵⁴. This multiprocessor clock frequency chart shows it almost leveled off in 2005 at 2.5-3.5 GHz with single core processor heat nearing the limit that heat sinks and fans could dissipate.



That's when multi-core processors came to market. From 2005 through 2013, greater core counts increased the overall work the processor could do, even though each core ran at lower speeds than Moore predicted. But by 2014, AMD and Intel once again focused on clock frequency to boost performance because of the 3-D transistor “heat relief”. Clock frequency again followed Moore's projection.

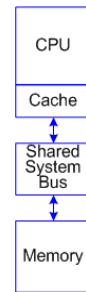


Aided by their 2007 work on an 80 core “Teraflops Research Chip”⁵⁵, Intel engineers were able by 2017 to resume the core race as the world saw an amazing 128 core, 3-D chip running at 32GHz. This collapsed the traditional CPU, cache, memory, and storage layers into one module, although secondary RAM and external storage tiers were still supported. Intense, global engineering work was underway over the ensuing three years and by the time GRACE was launched in 2020, a processor had 256 cores, one terabyte of cache memory, a quarter terabyte of dynamic memory, and a five petabyte solid-state disk (SSD).



Memory and Cache Management

A popular early microprocessor was the 1974 Intel 8080. Back then, the CPU spent a “reasonable” number of cycles accessing external memory, but soon processors became so fast that “memory wait states” caused significant execution delays. Cache memory helped reduce wait states. Smaller in capacity than main memory, cache contained frequently accessed data, and was significantly faster because of circuitry design and placement in or near the processor itself. When a processor needed data from main memory, it examined cache memory first, and if it was found, the “cache hit” data was retrieved extremely quickly. Otherwise, the “cache miss” data was found in main memory. Algorithms kept track of what was being accessed in memory and kept the most frequent data in the cache. The Intel 486 was the first of Intel’s family of processors to offer integrated cache on the chip.

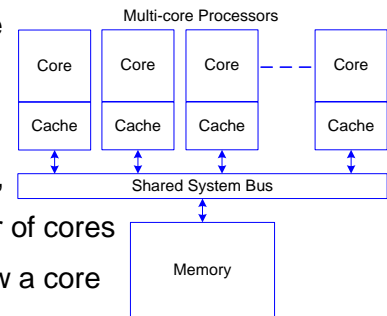


With the introduction of dual processor motherboards, advances in operating system design, and SMP, a server could *typically* do twice as much work as one with a single processor. Not all programs ran effectively with SMP designs, but “multi-threaded” applications took advantage of parallel execution code paths running on different processors, all under the control of a single operating system and shared main memory. The shared memory bus was an efficient way to access common memory with two (or a few) processors – this was called Uniform Memory Access (UMA). Intel Pentium and Pentium Pro were among the first to support SMP.

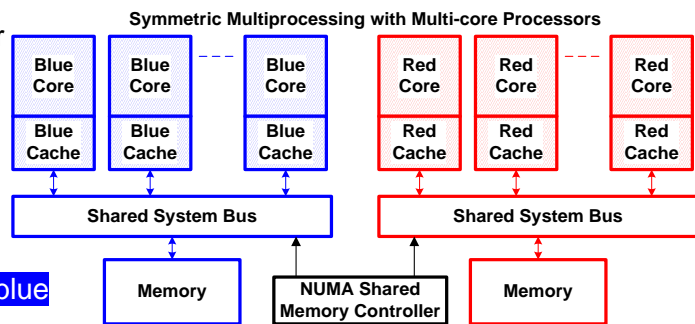
One of the issues that arose with UMA and SMP that had to be fixed before the design was finalized was *cache coherency*. Assume the **blue CPU** reads the value “123” from memory address 100 and puts it into its cache. Then the **red CPU** does the same thing. The **blue CPU** then changes the value of cached address 100 by writing the value “abc” in its place, which then

gets de-staged to main memory. What happens to the **red CPU** when it reads cached address 100 and gets the value “123”? The issue was called cache coherency and the problem was how to maintain the validity of each processors’ cache. This problem was fixed with “update” and “invalidation” – update (notify) other caches that memory address 100 has changed and supply the change so they were in sync, or invalidate which instructs other caches that address 100 is incorrect and they should get the correct value from main memory.

In 2005, multi-core processors were introduced allowing a chip to have two or more processing *cores*, each with their own cache, and all sharing main memory. A single two-core CPU could do the work of a dual CPU server at a lower cost, lower power profile, etc. For example, the AMD Phenom II X6 contained six processing cores. As the number of cores increased, UMA caused delays in accessing memory, which could slow a core down. Cache coherency became a bigger issue because of the number of cores that potentially waited while all the necessary update or invalidation messages were placed on the system bus.

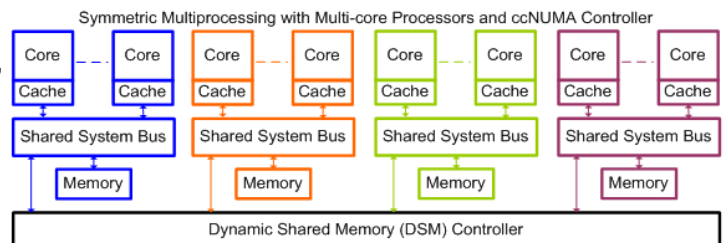


Internally, keeping track of which processor core was accessing a main memory location was a complex task – you can’t have two cores concurrently update the same memory location. Further, each processor had its own main memory, so a **blue core** application could access memory controlled



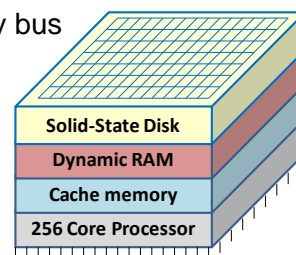
by the **red processor**. When this happened, the correct data was returned but with a slower access time compared to its own blue memory – i.e., asymmetric memory latency, or commonly called Non-Uniform Memory Access (NUMA). Some parts of memory could be accessed faster than other parts of memory, even when it was all on the same server. Both AMD’s Opteron and Intel’s Nehalem multi-core processors used NUMA memory access.

Cache Coherent Non-Uniform Memory Access, or ccNUMA for short, ensured each processor had accurate cache values. ccNUMA kept all cores’ cache and memory in sync. It allowed multi-core processors to scale without



enforcing difficult application and operating system reprogramming. With the contents of cache and memory in a consistent state, it was available to all the processing cores in a server.

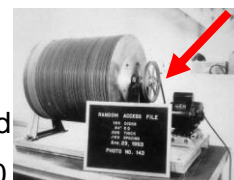
Given the traffic caused by the CPU needing data from memory, memory bus overhead, and latency, memory moved to a layer on the CPU chip. The processor, with 256 cores at layer 0, cache memory at layer 1, dynamic RAM at layer 2, and SSD at layer 3, faced the processor socket and started to resemble a chocolate layer cake. Each processor core supported 256 virtual machines.



With over a half-million virtual servers supported by a single 8-socket server blade, GRACE was reaching supercomputer status. Packaged with 144 blades per rack, and 160 racks, over 47 million cores each supported 10 virtual CPUs, all accessing their caches and main memories without a hiccup thanks in part to ccNUMA. Some might have sneered at this second coming of the mainframe, but most applauded the advance in general purpose supercomputing that brought about all sorts of efficiencies, economies of scale, and resiliency.

Storage

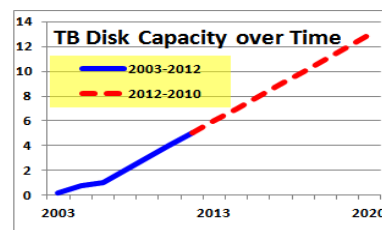
Mechanical hard drives, introduced in 1956, had come a long way. The first drive, IBM's RAMAC (Random Access Memory Accounting), weighed over 500 pounds⁵⁶. It held 5MB worth of data, equivalent to a single MP3 song, and used a belt driven motor. Rotating at 1,200 RPM, it had an access time of 800 milliseconds and retrieved a data record in about a second, or 2 to 3 blinks of an eye.



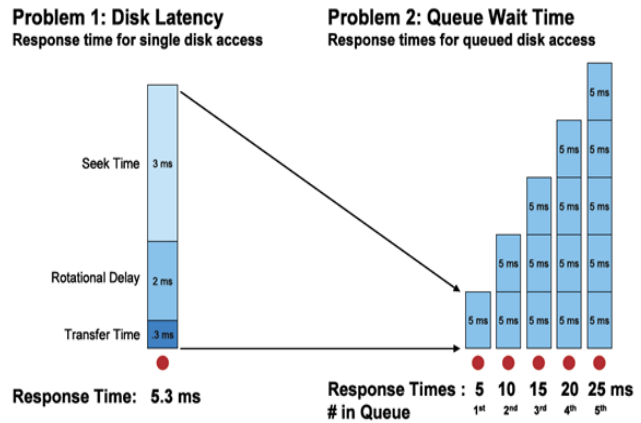
In 1970, IBM's 3330 drive using removable disk packs held 200MB⁵⁷. Each pack had a "cake cover" style removable plastic case, weighed 20 pounds, rotated at 3,600 RPM and had a 30 ms average access time. It wasn't until 1996 that a 10,000 RPM drive was available⁵⁸. Seagate's 9GB drive weighed about a pound and accessed a record in about 8 ms⁵⁹. In 2000, they shipped a 15,000 RPM 18GB unit with a 3.9 ms average access time⁶⁰.



Drive capacity reached 5TB by 2012 and 13TB by 2020, as areal density increases from 625 Gbsi to almost 1 Tbsi with the shift from perpendicular heads to Heat Assisted Magnetic Recording (HAMR) technology⁶¹.



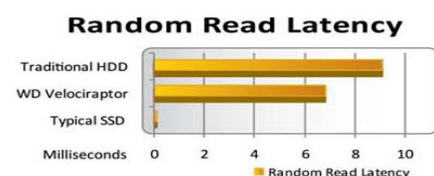
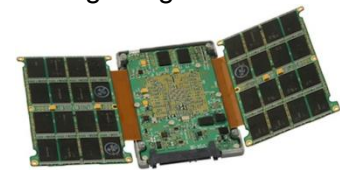
Unfortunately, drive performance did not keep up with drive capacity or CPU speeds⁶². Drives couldn't rotate faster than 15,000 RPM and rotation is key to its transaction rate. The more concurrent work you make a mechanical drive perform, the greater the I/O delay it causes the rest of the system.⁶³ For example, a 5.3 ms response time drive can quickly be overwhelmed as an incoming queue of I/O requests builds as shown to the right⁶⁴.



In an effort to increase performance, various mechanical disk techniques were employed in addition to command queuing. For example, striping data across multiple drives in a RAID group could yield a significant increase in IOPS⁶⁵. Short stroking took a large drive, like a 300GB unit, and only store data on the outer edges to minimize disk head movement. Thin provisioning, like RAID, sought to spread data across all the drives in a large pool of disks.

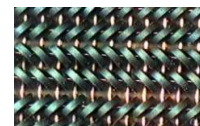
At the same time, drive form factors shrank from 3.5" in the 80's and 90's, to 2.5" and eventually leveled off at 1.8". Costs continued to drop every year. By 2015, the fastest drive still rotated at 15,000 RPM. Access times decreased from 3.4 to 2.9 ms and disks could handle 250 IOPS. Then, almost as a last gasp, drive manufacturers introduced dual actuator arms in their 2.5" drives, doubling its performance. By 2020, four actuators were commonly available on 13TB drives, achieving over 1,000 IOPS, due in part to the increased density of HAMR technology.

With data growing 60% a year and drives getting larger but not faster, storage frames resorted to new solutions in order to deliver fast response times. One significant change began in 2009-2010 as cost-effective, enterprise-quality SSD technology shattered the mechanical drives performance barrier, all without moving parts. SSDs produced less heat, and were so fast that they could randomly access data in microseconds whereas mechanical drives needed milliseconds⁶⁶. Mechanical drives needed to position the head to the right track (seek time) and wait for the platter to be rotated under the head (rotational delay) to read the data. All that went away with the SSD since the data was accessed directly – i.e., nothing rotates or moves in and out. SSD-equipped servers with multiple virtualized applications saw



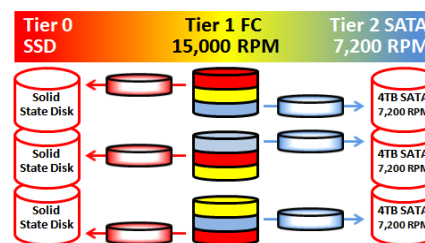
dramatic performance improvements. However, SSDs cost much more than comparable rotating disks, an issue that took years to be addressed.

The concepts behind SSD go back to the early 50's when computers used core and Charged Capacitor Read Only Store (CCROS) memory⁶⁷. In this image, you can see the “donut” cores where each bit corresponded to an individual core.



In 2009, hybrid strategies emerged incorporating SSD for applications needing low millisecond performance times with less demanding applications residing on slower mechanical drives of different speeds. For example, EMC announced their “Fully Automated Storage Tiering” promotion/demotion algorithm

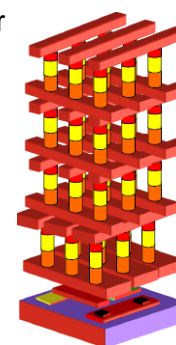
that moved data to a higher or lower performing SSD, Fibre-Channel, and SATA drive technology based on its profile and needs⁶⁸. With SSD in tier 0 and 4TB in tier 2 or 3, overall



performance increased while using fewer, larger capacity drives. Seagate’s Adaptive Memory Technology (AMT)⁶⁹ marketed a SSD/HDD hybrid drive that moved active data from a rotating HDD to SSD all within the same drive. Over time, tiering in the storage frame was reduced as SSD caught on and it was common to have SSD and 10TB SAS in a unit. On top of that, data was compressed, deduped, and encrypted, while unused data was archived to other platforms.

The SSD used non-volatile NAND (“not AND”) chips to store data. The chips use transistors to record their state as charged/erased or not charged/programmed – i.e., 1 or 0. In 2012, there was Single-Level Cell (SLC) and Multi-Level Cell (MLC) NAND memory. SLC retained the state of one bit while MLC retained the state of two or three bits. With fewer bits, SLC technology was faster. The two-bit MLC requires 4 voltage levels to retrieve 00₂, 01₂, 10₂, or 11₂. Likewise, three-bit MLCs need 8 voltage levels to be able to return 000₂ through 111₂, or 0-7.⁷⁰

The SLC write time was much faster than MLC’s while only being slightly faster with reads. The multi-state MLC was denser and costs less per bit stored. SLC lasted longer because silicon breaks down faster as it is charged more often⁷¹. MLC density reached 4 bits in 2006⁷², which means the cell can store 0-15. Using tri-gate transistors and 3-D circuitry, MLCs reached 8 bits per cell by 2017, or 0-255 stored in a single cell. When GRACE was introduced in 2020, the density reaches 64 bits per cell, or 2⁶⁴-1 (18,446,744,073,709,551,615 bits or 2,048 petabytes) – amazing innovation!



Not limited to just hard disk form factors, PCIe plug-in SSD cards became popular in 2008 allowing “disk storage” to be placed much closer to the server than SANs allow. It also meant that the data no longer had to be sent down an HBA to a SAN for the data to be retrieved, so I/O rates increased dramatically. This chart shows PCIe-based SSDs easily outperformed the fastest HDD by IOP factors of 4000:1 and the fastest SSD drive by 20:1. Fusion-io’s 2010 product offered an enormous 5TB per card and near instantaneous seek

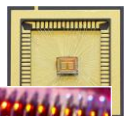
	Manufacturer	Model	RPM	Size	Interface	Read Seek Time	Avg Latency	Sustained Transfer Rate MB/sec	Power Watt	Max IOPS
HDD	Seagate	Cheetah 15K.7	15,000	300 GB	SAS 6Gb	3.4ms	2ms	122-204	12.9	200
SDD	Intel	320 Series	n/a	600 GB	SATA 6Gb	0.1ms	75-90µs	205-270	3.5	39,500
PCIe SSD	Fusion-io	ioDrive Octal	n/a	5 TB	PCIe x16	26µs	30µs	4,400-6,000	150	800,000

Notes: seagate.com/docs/pdf/datasheet/disc/ds_cheetah_15k_7.pdf
download.intel.com/design/flash/nand/325170.pdf
fusionio.com/load/-media-/1a7jn6/docsLibrary/FIO_DS_Octal_v16web.pdf



time, allowing server throughput to increase dramatically. They marketed a 20TB card in 2012. That year, EMC’s “Project Lightning” VFCache allowed their storage frame “smarts” to manage the data content of their SSD card, in essence, providing tier n-1 server read caching services⁷³.

Phase Change Memory (PCM) was another non-volatile memory technology rooted in the 60s. Marketed in 2016, PCM was useful for NAND MLC SSDs using the crystalline (low resistance or “0”) and amorphous (high resistive or “1”) states of chalcogenide glass to store data⁷⁴. Chalcogenide is used in rewritable CDs and DVDs. “Unlike NAND flash, PCM memory does not require that existing data be marked for deletion prior to new data being written to it – a process known as an erase-write cycle.



Erase-write cycles slow NAND flash performance and, over time, wear it out, giving it a lifespan that ranges from 5,000 to 10,000 write cycles in consumer products and up to 100,000 cycles in enterprise-class products. PCM could sustain 5 million write cycles...⁷⁵ and proved to be much faster than MLC NAND. By 2020, GRACE adopted PCM SSD for Tier 0 storage technology on 3-D processor chips and relegated MLC NAND-based PCIe SSD to tier 1 on the processor bus.

GRACE in More Depth – Securely Connecting with Any Device, Anywhere

Throughout the history of computing, various devices have been used for communications. Early on, interacting with giant number crunchers was through punched paper tape or cards. The computer communicated back with either pin-fed fan-folded paper or more paper tape (data storage). Teleprinters appeared in the early 60’s, but it was rare to enter a program or data directly into a computer.



Cathode ray tube (CRT) terminals came on the scene in the early 70’s. IBM’s 3270 and DEC’s VT52 were popular examples of TV-like devices that displayed what you were typing and

allowed you to make changes in the event of a typo. With the CRT, programs such as word processing, electronic mail, and office automation became popular as end-users began interacting with computers. Graphical terminals displayed bar, pie, and x-y coordinate graphs.

Personal computers were introduced in the late '70s with machines like the Apple I, TRS-80, and IBM PC 5150. Running a terminal emulation program, they could act as a DEC VT100. Computer mice, first conceived in 1963, became a popular way to interact with the machine. PC users accessed email when connected to CompuServe in 1979 using a telephone modem⁷⁶. The first full screen, ten pound, battery powered laptop was the 1984 Data General One. In 1993, the Mosaic Internet browser got the world hooked on the Internet. These devices and applications set the stage for cloud technology in 2008.



The PC era ushered in Windows and MAC graphical desktops in the mid-80's. Touch screens, a concept dating back to the 60's, made a resurgence with smart phones and tablet computers. Even so, the desktop was linked to the underlying hardware. By 2013, virtual touch screen cloud desktops emerged. Independent of local devices, they also masked the underlying remote hardware. Similar to graphical thin client terminals of the 90's, the virtual cloud desktop was a seamless local gateway to the cloud where a user opened remote files, programs, and even games. Cloud Desktop Operating Systems (CDOS) shown below from Cloudo⁷⁷, Jolicloud⁷⁸, and eyeOS⁷⁹ presented customizable desktops that emulated the users' physical desktop. These CDOS were later based on Amazon's Silk split browser⁸⁰ which offloaded intensive client operations to the cloud for efficiency. By 2020, the graphical environment was so seamless that users were unaware of whether they were using a private, public, or hybrid GRACE system or even where the data resided. Their metadata preferences decided where it ran.



The widespread popularity of Palm and BlackBerry smart phones began in the early 2000's, and flourished with Apple's iPhone in 2007⁸¹. Not just a phone, it provided access to multimedia, email, calendars, texting, e-payment, language translation, barcodes, and cloud computing. In many cases, they had more compute power than PCs from a decade earlier. By 2011,



“...billions of people worldwide don’t have computers but they do have smart phones.”⁸² These devices also generated a significant amount of digital data crumbs, which of course were also stored in the cloud for subsequent “big data” analysis by marketing companies and others.

These mobile devices created a huge amount of redundant network traffic, especially in public venues. For example, at a baseball game thousands of fans searched for the same data as they exited the park – team standings, rivals’ scores, and traffic conditions. “Smart” GRACE networks were installed and robotic blimps floated overhead at these public places to cache data transactions and service it locally, rather than straining GRACE resources. In metropolitan areas like New York City’s Central Park, the latest weather, transit schedules, movie times, and other common information was grouped by GPS location and cached so the weather report you received was the same one someone else asked for just a few seconds ago, all of which originated from a single GRACE update 5 minutes earlier. These caching proxy techniques saved hundreds of millions of duplicate transactions a year in densely populated areas.



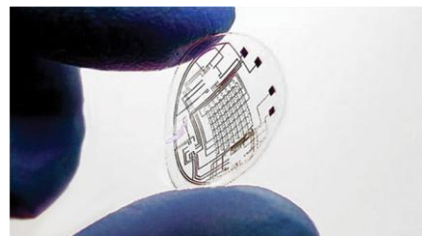
Even digital televisions (computers) became cloud-enabled through Video On-Demand (VOD). “...VOD services are now available in all parts of the United States. In 2010, 80% of American Internet users had watched video online⁸³.” Apple also introduced their sexy iPad that year, and although it lacked a mechanical keyboard, mouse, modem, and a printer, it offered access to thousands of applications and made cloud computing hip. This technology became so disruptive to traditional IT that by 2010, Apple’s Steve Jobs announced the end of the PC era⁸⁴. The world had taken a major step away from hard-wired connections to the mainframe, minicomputer, or other traditional computer resources and now their “computers” depended on the cloud. “In 2000, the total Internet traffic was just over 1 exabyte, and in 2010 will be about 256 exabytes, corresponding to an annual growth rate of 70% over ten years.”⁸⁵

While storing tablet music, photos, and other data in the cloud was the rage in 2013, digital paper became the “must have” Christmas present in 2014, dethroning the tablet. No longer burdened by a rigid device, it could be rolled and folded, and everyone carried a piece. Touch sensitive like the iPad, but weighing less than a milligram, it was used everywhere and leveraged the new LED-wireless systems that were just becoming available (all light sources attached with a wire to the power grid could act as a light spectrum router to the Internet). They were so popular that restaurants



began using them instead of printed menus so patrons could actually “see” the dish they wanted to order.

Based on Dr. John Rogers’ work at MC10⁸⁶, it was fashionable to wear digital contact lenses in 2017 as stretchable, bendable, ultra-thin electronic circuits became commonplace. Leveraging the brain’s compute power and the body’s electrochemical power, these disposable devices offered a “heads up display”



that bio-electrically integrated with GRACE when placed on the eye. With lightning-fast cloud response times, they all but doomed the latest 5G cellular communications efforts of Verizon and AT&T. Using touch, smell, hearing, taste, and sight, people seamlessly used GRACE in 2020 to enhance their daily lives, breaking the traditional user interface computing barrier.

GRACE in More Depth – The Labor Force

The data explosion was not new. The NY Times in 1967 published “Data ‘Explosion’ For Government”⁸⁷ and said “Almost no phase of a citizen’s life, from birth through income taxes to death, goes unrecorded by the electronic calculator in today’s society”. Data growth was real back then, but they didn’t envision the need for skilled data workers. Data increased 50-60% a year, brought on in part by the abandonment of analog devices and paper reporting. As more information was distributed through the Internet – current events, digital periodicals and books, broadcast and pre-recorded media – more data “crumbs” were created to feed the insatiable appetite of the advertising world as they demanded more access to consumers. As users accessed content through mobile devices, more GPS-encoded data was added.

DATA ‘EXPLOSION’ FOR GOVERNMENT

Electronic Calculators Keep Track of Moon and Mail

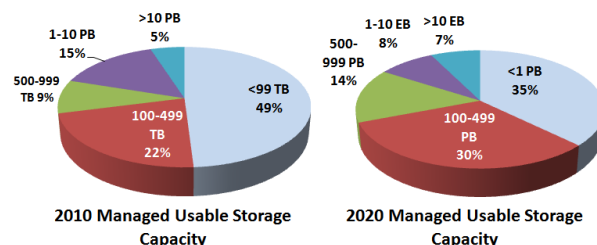
By EVERT CLARK
Special to The New Times
WASHINGTON—Once every minute on the average a car is stolen in the United States and a Federal Bureau of Investigation computer here is told about it.

A California highway engineer seeks an alternate way to route freeway traffic. He asks a computer in Sacramento, the state capital.

In Baltimore, a device called an optical page reader scans the quarterly reports of employe earnings, filed with the Social Security Administration—35 million industrial companies—needs what it sees

Government, being the largest of all large institutions in the complex modern world, makes perhaps the greatest use of computers. Almost no phase of a citizen's life, from birth through income taxes to death, goes unrecorded by the electronic calculator in today's society.

Everything was getting smaller and smarter as software was embedded in almost every product. Appliances, cars, and cereal boxes became intelligent devices. IDC said “We always knew it was big – in 2010 cracking the zettabyte barrier.

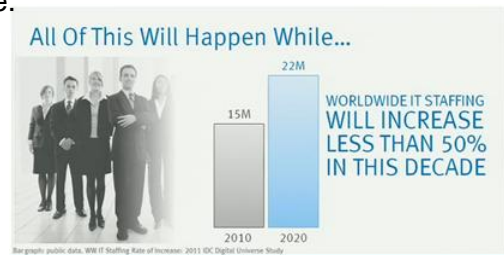


In 2011, the amount of information created and replicated will surpass 1.8 zettabytes (1.8 trillion gigabytes) – growing by a factor of 9 in just five years...while the pool of IT staff available to

manage them will grow only slightly.”⁸⁸. There was 200 times more data in 2020 than in 2012, businesses kept more, and fortunately storage costs dropped by a factor of 1000⁸⁹.

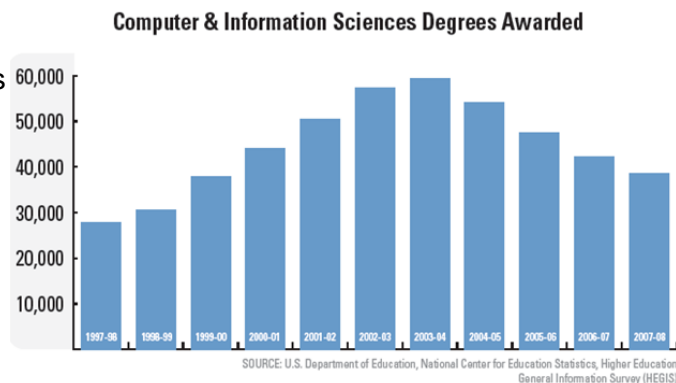
Numerous world recessions coupled with the productivity gains of advanced management software caused the ratio of IT “workers per TB” of storage to decline rapidly. Two years before Joe Tucci retired, the President and CEO of EMC Corporation was quoted as saying “...the data deluge will grow 44 times larger by the end of this decade.

About 90 percent of that information being created is unstructured. In 2011, the digital universe contains 300 quadrillion files to manage. Information is growing at a phenomenal rate, yet IT staffs will grow by less than 50 percent this decade.”⁹⁰

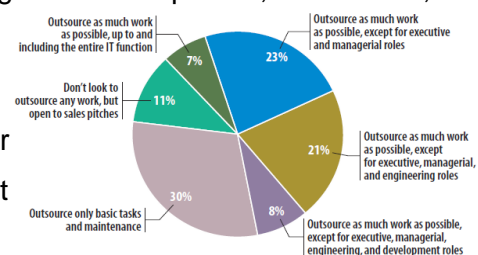


In the United States, the decline in the IT labor force began around the height of the Y2K crisis as Baby Boomers of the late 40’s began to retire. Through 2010, fewer and fewer students graduated with computer science degrees.

College students decided against IT careers because of a desire for job security and the threat posed by outsourcing. Meanwhile, employers complained that they couldn’t hire enough skilled local IT workers (perhaps at the right price).



With the pool of trained U.S. IT workers shrinking and the employers’ desire to spend less on IT, outsourcing became popular in the 21st century. Outsourcing relies on expertise, lower costs, and the ability to speak the local language. One survey showed almost 60% of executives were “outsourcing as much work as possible”⁹¹. A remote IT “virtual” administrator using data communications performed business functions at a lower cost than local staff.



For U.S. IT workers, outsourcing “...sparked widespread debate and a political firestorm three years ago, it has been portrayed as the killer of good-paying American jobs. ‘Benedict Arnold CEOs’ hire software engineers, computer help staff, and credit card bill collectors to exploit the low wages of poor nations. U.S. workers suddenly face a grave new threat, with even highly

educated tech and service professionals having to compete against legions of hungry college grads in India, China, and the Philippines willing to work twice as hard for one-fifth the pay. Workers' fears have some grounding in fact. The prime motive of most corporate accountants jumping on the offshoring bandwagon has been to take advantage of such "labor arbitrage" – the huge wage gap between industrialized and developing nations. And without doubt, big layoffs often accompany big outsourcing deals.”⁹²

The movement to cloud computing seemed to break the back of traditional IT outsourcing in late 2012. While software skills were still in demand, especially with the surge of smart devices accessing the cloud, the community added new “cloud” jobs such as cloud architect and cloud engineer. Titles continued to evolve benefiting from the move to a dynamic model that focused on data science. The number of long-term outsourcing contracts also tailed off as businesses saw the benefits of software-as-a-service and spent their resources on core activities. There was a shift away from private to the public cloud, although this took another decade or two to complete due to the time it took to re-platform legacy applications.

Meanwhile, the number of software and hardware manufacturers continued to decline through mergers and acquisitions, and those that survived focused on cloud migration services. IT workers who feared the cloud would put them out of a job, especially as the number of data centers decreased while hardware reliability increased, found retraining in analysis of “big data” to be a new, rewarding field of expertise.

By 2014, the CAPEX savings promised by cloud computing became a reality and corporate IT costs finally came down. The IT workforce increased somewhat – just enough to tackle the IT backlog that had grown during the recessions. Those jobs, previously outsourced for skill set and monetary reasons (i.e., high health care costs), returned to the United States. Expertise grew in low-cost but highly educated areas such as North Dakota and Utah. The flexibility of software-as-a-service actually favored domestic, low-cost locations. High-speed Internet access helped smart companies tap into a rural, inexpensive yet stable labor force providing back-office and call center activities from small delivery centers and home-based staff. The cloud and distributed applications accelerated this trend and helped put people in the U.S. back to work again.

By 2020, the consumer’s detritus data was used by popular “big data” advertisers worldwide like Catalina Marketing⁹³. By instantly data mining GPS-enabled smart phones and frequent

shopper card devices, Catalina could display an e-sign for “Mrs. Smith” as she passed a grocery window on her way home from work that said “Mrs. Smith, your son Billy would love some Oreo cookies. You haven’t bought them in a long time and they are on sale today”.



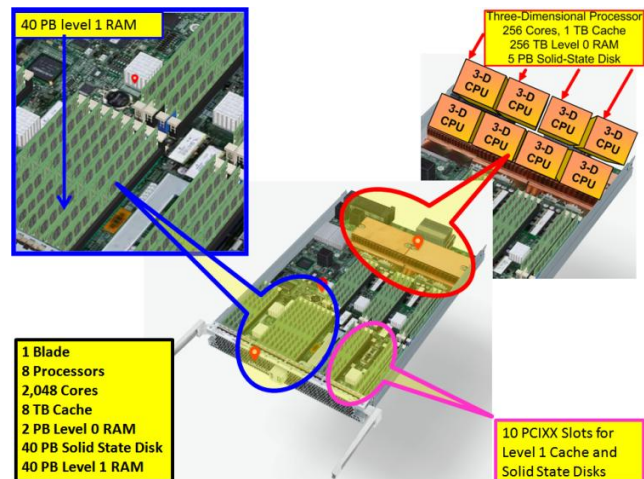
As GRACE became more popular in the 2020s, the lines that separated traditional computing from private-public-hybrid clouds blurred. Workers shifted between traditional IT positions and cloud outsourcing. Self-service IT morphed into Automated Information Technology, following in the footsteps of Edison’s and Tesla’s utility electric power generation model.

GRACE in More Depth – Form Factor

GRACE’s building block was the 3-D processor, a low-heat, high density four-level design integrating 256 processor cores, 1 TB of cache, 256 TB of Level 0 RAM, and 5 PB of SSD.

The 3-D processors were socketed on 4 inch wide computing blades along with RAM and

PCIeX expansion slots for cache and SSD modules, and placed in low-profile, high-density blade servers. PCIeX, a fifth generation 128 bits-wide PCI interface provided 128 GB/s of throughput and was 1,000X faster than 2013 PCIe 4th generation interfaces. A blade held 8 processors (2,048 cores), 8 TB of cache memory, 40 PB of level 1 auto-tiering RAM, and 40 PB of auto-tiering SSD. This cost-effective packaging was



energy- efficient and powerful because it was built from extremely low-latency components and used smaller electrical paths.

Blade servers have multiple benefits in the GRACE architecture:

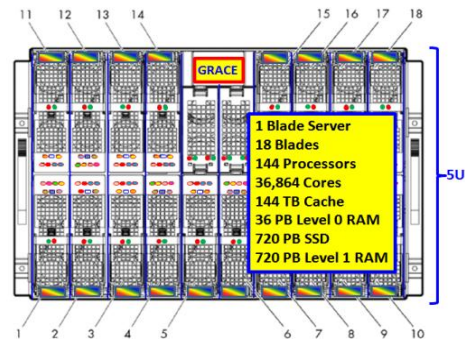
1. Reduced energy costs – housing 18 blades per blade server reduced power requirements. Sensors on the blades also allowed them to be turned off under GRACE virtualization control when not in use. Power was supplied using redundant copper strips fed from alternate power circuits from the rear of the rack.
2. Reduced cabling costs and complexity – the blades and servers leveraged an etched matrix backplane to eliminate complex wiring.



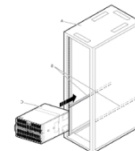
Blades were inserted or unplugged from the server on demand to aid serviceability. The server could also be non-disruptively repaired once GRACE ejected it from the backplane. The top of rack (TOR) 1U communications module supplied the 1,000 Gb Fibre Channel over Ethernet (FCoE) connections.

3. Reduced CAPEX – ultra-optimized, standardized blades significantly lowered acquisition costs compared to individual servers. As technology evolved beyond 2020, newer, more powerful blades and blade servers replaced older ones. Blades and servers of different vintages could also co-exist to protect hardware investments.

Eighteen blades were housed in a 5U-tall blade server chassis. A chassis containing over 36,000 processor cores, all supported by auto-tiering RAM and SSD. Auto-tiering allowed the most actively used RAM and disk data to rise or fall to the most efficient level based on usage patterns, similar to EMC's FAST in their 2010 high-end storage arrays.

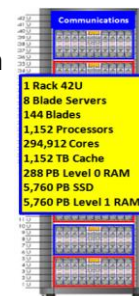


Eight blade servers and dual 1U communication servers fit into a 42U rack. The rack had a copper etched backplane that simultaneously supplied power and data communications using independent, redundant paths for information and management coordination. The rack vented heat from the top using powered exhaust fans to prevent hot/cold data center aisles. GRACE monitored each server and blade, and automatically moved workloads off any device due to high-heat, underperformance, or other issues. Weighing 32 pounds, servers were serviced by a single staffer. A step-stool was supplied for servers near the top of the rack.

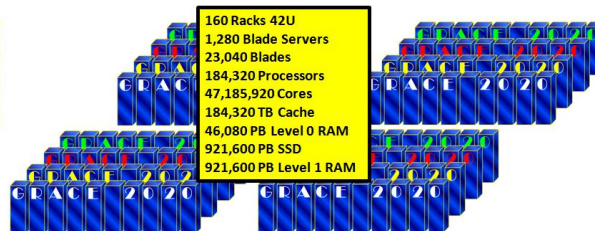
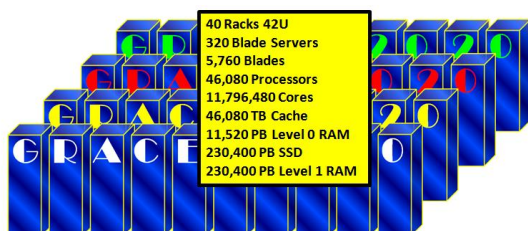
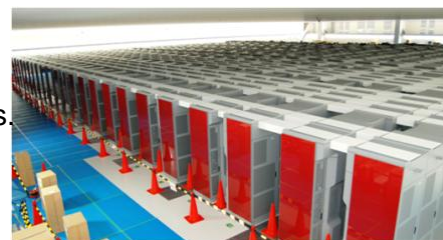


The TOR communications server connected the rack with other racks using a high-speed 10 TbE Rack Connection Service (RCS) for ultra low inter-rack latency. The TOR RCS Ethernet interconnect used a 2011 Arista Networks leaf-spine design to support thousands of servers.⁹⁴

A rack held 8 servers, 144 blades, and almost 300,000 processor cores. Through compression and intra-GRACE deduplication, backing up and restoring files was easy and efficient as the managed backup/disaster recovery data flowed through the backplane and RCS to a lower tier of storage at another GRACE 2020 site.



Depending on the GRACE location, various rack packaging permutations were used. In low Internet traffic areas, 40 racks were often brought together as a cohesive public cloud servicing millions of users and billions of transactions per second. Racks were interconnected through dual TOR RCS grids and allowed rapid, non-disruptive connect/disconnect of racks. The solution was highly elastic and additional capacity could be brought online within 24 hours. For larger metropolitan areas, the system could scale to over 47 million cores and nearly an exabyte of storage.



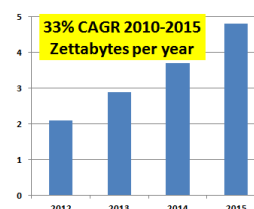
A business could choose the private or hybrid GRACE 2020 unit they needed. Data scientists speculated the design could be further expanded another 10-fold based on the latest Moore's Law calculations and the fact that all GRACE implementations were united in a worldwide grid.

GRACE	per Blade	per Server	per Rack	40 Racks	160 Racks
Blades		18	144	5,760	23,040
Processors	8	144	1,152	46,080	184,320
Cores	2,048	36,864	294,912	11,796,480	47,185,920
TB cache	8	144	1,152	46,080	184,320
PB RAM Level 0	2	36	288	11,520	46,080
PB SSD	40	720	5,760	230,400	921,600
PB RAM Level 1	40	720	5,760	230,400	921,600

This was just the beginning of GRACE. As the immense power and efficiency of the public version spread the land, companies that had stubbornly supported traditional data centers and private GRACE implementations began to flock to the public version. All their data processing needs were being serviced in a manner similar to plugging a lamp into an electrical outlet. They returned to running their core business without the data center burdens they used to have.

The End of Cell Towers?

As the cloud caught on, Wi-Fi and broadband cellular communication workloads increased. Through 2011, laptops, tablets, smart phones, and other devices relied on their built-in radio receiver/transmitter to log in to wireless systems for half-duplex communications. In 2012, Rice University researchers came up with simultaneous bidirectional data transmissions on the same frequency⁹⁵. Full-duplex Wi-Fi doubled wireless speeds, but it wasn't a commercial success until 5G networks



rolled out in 2015. Network traffic increased to “...4.8 zettabytes per year by 2015 or every man, woman and child watching a full length movie once a day for one year.”^{96 97}

Cellular service relied on a system to hand off mobile calls as users ventured beyond the communications limit of one cell to another. First generation cell phone towers appeared in the late 70's⁹⁸ with fairly ugly directional antennas on top. Antennas were aimed at other antennas located on other towers or on tall buildings.



Interestingly, these towers were not laid out for circular coverage but rather with hexagonal coverage to avoid gaps that would have arisen with circular designs⁹⁹.



With every generation of service introduced, the distance spanned by these hexagons was forced to get smaller and smaller – i.e., more ugly towers than ever as the speed increased!

As more cell phones were sold, more towers were needed to handle the data traffic. By 2012, the U.S. population reached 315 million, and fourth-generation cellular service was rapidly rolled out to support 323 million

CTIA -US	Jun-96	Jun-01	Jun-06	Jun-11
Wireless Subscribers	38M	118M	220M	323M
Wireless Penetration	14%	41%	73%	102%

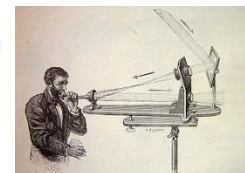
subscribers. While not everyone had a cellular device, others had multiple devices¹⁰⁰ and micro cell towers helped carry the workload. Smart phones strained cell networks and customers' bills. For example, playing a 200 MB Netflix video on an iPhone used 1/10th of a customer's monthly 2 GB data allowance¹⁰¹. Wireless carriers began placing limits on customers' usage for fear of over-taxing their network.

As 2016 neared, the average person had three wireless devices and 5G networks¹⁰² barely kept pace with bandwidth demands. Local WiFi speeds exceeded

Technology	1G	2G	3G	4G	5G
Start	1970	1980	1990	2000	2010
Deployment	1984	1995	2002	2010	2015
Bandwidth	2K	4 - 64K	2M	.2 - 1G	>1G

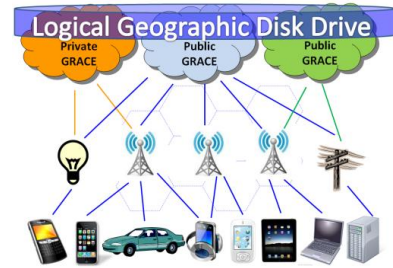
7Gbps using the 2012 WiGig 60GHz design¹⁰³. The number of cell towers in America grew 11X from 22,663 in 1995 to 253,086 in 2010, or one per 15 square miles. By 2020, 4G and 5G were both in use on almost 400,000 towers, or one per 10 square miles. Towers were expensive and moving to 5G meant they had to be closer together than with 4G.

Cell towers were greatly oversubscribed, and lost calls and dropped data packets became a common occurrence. Fortunately, wireless devices were also connecting with a new technology – visible light communications. Alexander Graham Bell in 1880 invented a photophone that used light to send data¹⁰⁴ and Harald Haas found a way to use LED lights in Li-Fi to send information in 2004¹⁰⁵. While Wi-Fi, broadband and Li-Fi were part of the electromagnetic frequency spectrum, Li-Fi did not use radio waves.



By 2020, light fixtures used Li-Fi enabled “light bulbs” with tiny receiver/transmitter circuits that carried data traffic worldwide on existing power lines. In the end, a smart phone or tablet could access GRACE using the newly enhanced 10 GHz graphene¹⁰⁶ Wi-Fi and cellular networks, or with Li-Fi by merely passing a lit storefront, streetlight, or a car’s headlights at night.

Called “PowerLine Carrier”, the concept of using electrical wiring as data cables was invented in 1928 by Bell Laboratories¹⁰⁷. In offices, data cables were quickly becoming a thing of the past as Ethernet data could be sent through the radio spectrum, overhead florescent lights, or any electric outlet. Data sent over distances could traverse the national electrical grid. Cars used towers to communicate telemetry, receive entertainment, directions, and monitor the health and safety of its passengers. At homes, the expense of wiring the “last mile” was eliminated.



GRACE Removes Barriers to 21st Century Cloud Computing

Cloud computing had many detractors. Issues plagued it for years. Initially, companies such as Amazon and Google hyped the technology, especially when the former introduced their supercomputer in late 2011¹⁰⁸, but collectively they had a hard time convincing corporations to run their IT operations in their cloud. Security, reliability, controls, standards, and compliance were among the top concerns. Companies were afraid their data could be stolen, downtime could be crippling, the lack of multi-tenant oversight could be embarrassing, the proprietary nature of each cloud vendor, and data locality could subject it to laws of another state or country.

There was a push towards private cloud computing in 2010 where lower costs, self-provisioning, on-demand use, scalability, consolidation, and other benefits became significant. Individual users and other organizations with less demanding requirements flourished in the public cloud. They saved money, removed maintenance headaches, and enjoyed all the benefits of having a vast computational environment at their fingertips. So it was natural that hybrid computing which married the two concepts seemed to be the ideal solution – total privacy for those who needed it and total freedom for others, plus the ability to float between the two clouds.


GRACE’s flexible metadata policies and ironclad security model guaranteed safe and flexible computing without the negatives originally associated with the cloud. Not only were the issues of security, reliability, controls, standards, and compliance addressed, but applications remained

as private as its DNA signature. Automatic failover prevented the loss of a transaction. Every organization utilized the reporting function and maintained data oversight. End-users tweaked extensive metadata controls. A uniform set of APIs allowed for seamless data interoperation.

In addition, GRACE provided single-instance storage, eradicated viruses, banned hackers, and cleaned up all the nastiness of traditional computing. GRACE-to-GRACE disaster recovery was a popular selection when natural or manmade calamities occurred. Full backup and restore, and automatic archival of older data was also available. Momentum increased as corporations moved applications rooted in their traditional data centers into the three GRACE cloud models.

Vendor Lock-In, Pricing Models, and Service Level Agreements

A fundamental stumbling block for early cloud computing endeavors was the fear of vendor lock-in. Proprietary solutions meant trying to move or port your application and data from one environment to another, be it a cloud or non-cloud environment. As a risk-filled, expensive, and often time-consuming adventure, it was to be avoided almost at all costs.

By establishing a set of standards and insuring interoperability through common platforms and APIs, applications developed in the GRACE framework were guaranteed to work on any private, public, or hybrid GRACE. Standardization was key to the computing experience, similar to how McDonald's hamburgers taste the same whether they were consumed in Tokyo  or New York City.

On top of that, cloud computing seemed open-ended given the complex billing formulas used by early hosting data centers. For example, Amazon's 2011 rates for U.S. East users was different than U.S. West, Asia Pacific (Tokyo), Asia Pacific (Singapore), or EU rates, making workload



budgeting difficult a priori if they automatically shifted locales. Think about a taxi ride to a destination you've never been to before – it is

hard to estimate the cost based on the sign on the cab's door.

GRACE's standardized pricing was introduced when the major telephone companies became its principle owners. Years of experience with cable TV and monthly billing allowed them to seamlessly merge communications and computing billing for users and companies. With flat-

Amazon Web Services Pricing	US East (Virginia)		US West (Northern California)	
	Linux Usage/hr.	Windows Usage/hr.	Linux Usage/hr.	Windows Usage/hr.
Standard On-Demand Instances				
Small (Default)	\$0.09	\$0.12	\$0.10	\$0.13
Large	\$0.34	\$0.48	\$0.38	\$0.52
Extra Large	\$0.68	\$0.96	\$0.76	\$1.04
Micro On-Demand Instances				
Micro	\$0.02	\$0.03	\$0.03	\$0.04
Hi-Memory On-Demand Instances				
Extra Large	\$0.50	\$0.62	\$0.57	\$0.69
Double Extra Large	\$1.00	\$1.24	\$1.14	\$1.38
Quadruple Extra Large	\$2.00	\$2.48	\$2.28	\$2.76
Hi-CPU On-Demand Instances				
Medium	\$0.17	\$0.29	\$0.19	\$0.31
Extra Large	\$0.68	\$1.16	\$0.76	\$1.24

* Windows® is not currently available for Cluster Compute or Cluster GPU Instances
 ** Cluster Compute and Cluster GPU Instances are currently only available in the US East Region.
<http://aws.amazon.com/ec2/pricing>

rate plans, customized usage schemes, and subscription contracts, costs were more predictable than those of a decade earlier. Budget-based pricing could even dynamically set application resources so it would run on a small memory sub-core to stay within a budget, and so forth.

Each GRACE licensee posted an assurance bond under government control to guarantee their GRACE cloud would be funded for a period of 24 months in the unlikely event they failed as a corporate entity. Auditable through a SAS 70¹⁰⁹ accounting firm, violations of GRACE policies could result in a loss of a GRACE license to operate and a forfeiture of its bond.

Another major issue that GRACE tackled was service level agreements (SLAs). Throughout the history of hosting, the Internet and, cloud computing, ensuring and pricing SLAs to match computing needs was left to sharp negotiators. When the service level was not met, the customer often had little recourse but to threaten the provider because of the pain of moving applications and data to another vendor. The time it took to act could easily jeopardize the customers' business and the providers' issues could be complex, expensive, and require patience, all the while, the clock could be ticking. Legal action could result in a cash reward, but it could come too late if the customers' business lay in ruins. For example, Amazons' Elastic Compute Cloud (EC2) had a 99.95% annual uptime guarantee that provided service credits against future bills for substandard service¹¹⁰. If you used EC2 and the Simple Queue Service or Simple Storage Service which did not have SLAs, you might not qualify for the credit.

GRACE employed a fluid application/data model that leveraged standardization. If one provider did not meet your expectations, you could near-instantly take your business elsewhere. Highly elastic and scalable, any small hiccup in service levels would be flagged by advanced proactive monitoring software that shifted running workloads to alternate gear in real-time. It was not a non-stop environment, but it was as close to it as technology and economics would allow.

Where were GRACEs Located?

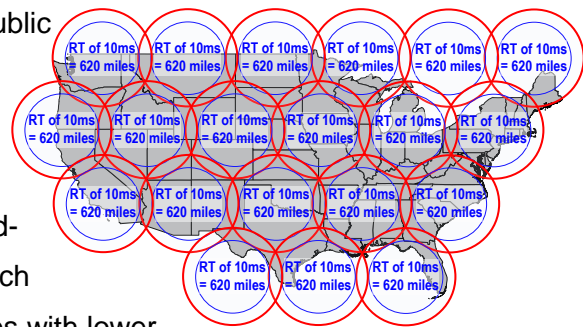
In 2011, "...a quarter of the world's population—1.5 billion people—connect to the Internet. Over the next five years, there will be over 10 billion client devices and another billion users accessing hundreds of thousands of Internet services, all competing for limited IT resources that will likely choke traditional infrastructures without a radically simpler, more secure, and more efficient way of doing things."¹¹¹

By 2018, GRACE was ready to take this challenge head-on. Data scientists developed the GRACE Data Center (GDC) model and tackled the issue of where to place them. Relying on

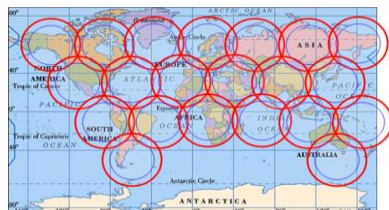
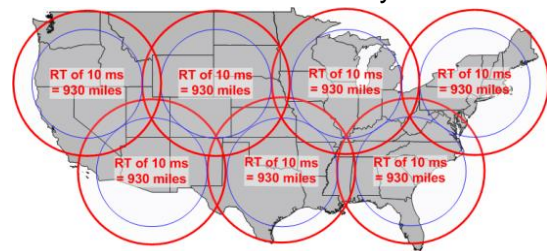
Fujitsu 50 Tbps optical interconnects from 2011¹¹², GDCs offered low transaction response times. The issue was the time it took a light wave to travel. The speed of light (SOL) in a vacuum is 186,000 miles/sec, but when sent down a hair-thin glass tube, refraction issues degraded the SOL by 1/3 to 125 miles/ms. To get data back and forth between GDC's in less than 10 ms, they could be no more than 623 miles apart without taking into account other networking delays and computing overhead.

Speed of Light	miles per sec	miles per ms	RT miles per ms	RT miles per 10 ms
Vacuum	186,000	186	93	930
Fiber	124,620	125	62	623

To address transactions that spanned two or more public GDCs, intelligent buffer and optimization devices employed deduplication, compression, out of order packet delivery, and other techniques to deal with latency and packet loss. That meant a 620 mile round-trip design in the U.S. needed 20 GDCs. This approach made it easy to automatically shift workloads to places with lower kWh electrical costs. True application and data mobility followed the user to wherever they traveled. Universal email boxes "@GRACE.COM" followed the GRACE user as they moved around the world.



As GRACE matured in 2030, the fiber optic refraction penalty reduced and the same coverage took just 12 public GDCs while still delivering a 10 ms maximum



round-trip latency. There were still private GRACE deployments, although as time went on, there were fewer of them worldwide. By 2040, GRACE was well into it's third revision with even larger spheres of worldwide computing.

GRACE's Metadata Dashboard

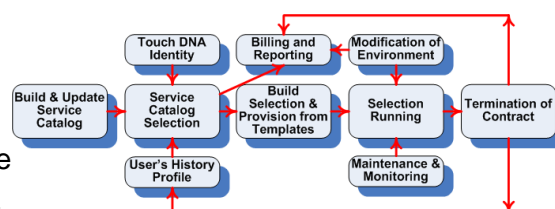
GRACE's metadata and graphical controls focused on the customer experience, the business functions placed into the GDC, and its own technical teams to ensure that operations ran without interruption. Strict enforcement of metadata controls assured users that any programs in the service catalog had full GRACE permissions as vetted by automated scanning procedures. Those assurances promoted safe and trusted computing for all, promoting digital peace of mind, similar to an electrical appliance having the Underwriters Laboratories logo.



The end-user experience began when they created a GRACE profile. Credentials asserting their baseline identity were confirmed in advance by the non-intrusive touch DNA registration process to prevent criminals from accessing GRACE. Algorithms examined their cloud computing behavioral patterns and when judged to be responsible and bona fide users that would abide by GRACE's guidelines, their credentials were safely and securely stored immutably in GRACE's metadata repository. From there, users selected their computing environment, choosing applications and databases, middleware software, file systems, etc. as called out by a service catalog options list, similar to ordering a customized cup of Starbucks coffee.



In more detail, the end- or corporate-user applied for GRACE services from the self-service metadata GUI. Starting with a touch DNA ID, they entered a credit card for billing, and chose a program or utility from the filtered service catalog based on their profile, such as



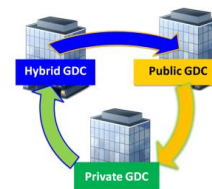
a cloud desktop with word processing, spreadsheet, and photo editing. If they were a corporate user and permitted access to a specific business program or environment based on job function, that view was added to their catalog. A corporate staffer could also set up a specific batch schedule of jobs to process business operations.

Tunable metadata adjusted each program's virtual environment through radio-button settings. Choices could be made for memory, performance, deduplication, compression, and encryption. They selected hourly, daily, or weekly backups, and

Extensible Metadata		Projected hourly cost of options selected \$ 1.29				Notes
User Profile	DNA ID [.....touch.....]					Touch DNA
Application	name *****	email contact *****@*****	phone contact (xxx) xxx-xxxx	credit card # *****		Application name, contact and billing info
Organization	unique id *****					Unique organization ID if applicable
Single sign-on	yes <input type="radio"/>	no <input type="radio"/>				Preference
Performance	slow \$ <input type="radio"/>	medium \$\$ <input type="radio"/>	fast \$\$\$ <input type="radio"/>	n/c \$ <input type="radio"/>		Up to 1, 2 or 4 Virtual Processors
Memory	slow \$ <input type="radio"/>	medium \$\$ <input type="radio"/>	fast \$\$\$ <input type="radio"/>	n/c \$ <input type="radio"/>		Up to 100GB, 500GB, 1TB
Disk	slow \$ <input type="radio"/>	medium \$\$ <input type="radio"/>	fast \$\$\$ <input type="radio"/>	n/c \$ <input type="radio"/>		SSD Tier 0, SSD Tier 1, Rotating
	yes \$ <input type="radio"/>	no \$\$ <input type="radio"/>		n/c \$ <input type="radio"/>		

whether cluster or disaster recovery protection was needed. To aid the courts, legal data jurisdiction and compliance could be set as well as whether the user wanted discounted or even free GRACE use by selling the rights of their work to advertisers. Choices dictated the projected hourly cost. Optional "no choice n/c" settings allowed GRACE to choose the best service level based on not-to-exceed pricing and market conditions. Once complete, they were ready to go.

GRACE was easy to customize. For example, a military contractor with a "big data" mid-air refueling system could configure a large memory model



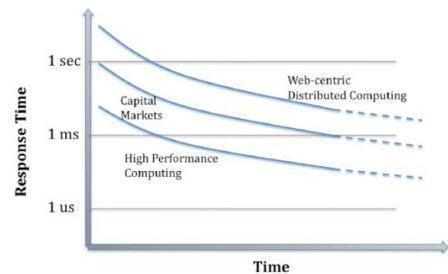
with 4 virtual processors, tier 0 SSD, and a near real-time priority. It cost more than a setting of default memory, 1 virtual processor, basic disk, and low priority. Commodities like disk space were billed based on preferences such as compression, deduplication, encryption, archiving, tiering, and location.

There were GRACE model choices and whether a program could move real-time between GDCs – i.e., start in a private NYC GDC, move to a London hybrid GDC after hours to lower costs, and shift back to a public NYC GDC for weekend processing. Month-end choices could move the workload to a New Jersey public GDC for higher performance. Based on choices, the user could get an estimated bill or opt for fixed price billing. This approach made for predictable operational expenses and its uniformity made price comparisons straightforward. With final approval, the application environment was built from templates and within minutes, the service was ready to use.

With the movement to the cloud, GRACE’s customers wanted a blazingly fast and inexpensive experience. They almost took for granted that it was secure and environmentally green. Dashboards showed the workload cost as well as year-to-date and weekly views permitting users to make decisions on runtime priority, where the environment should run, disk tiering, or over a dozen tunable OPEX parameters. They will also want to see performance metrics on how well things were going.



Administrators could tune their customers’ web experience, tracking things like response times to product catalog page turns and adjusting memory, processing, response time by industry profile, and disk tier metadata¹¹³. SLA violations and program sequence wall clock times were also available.



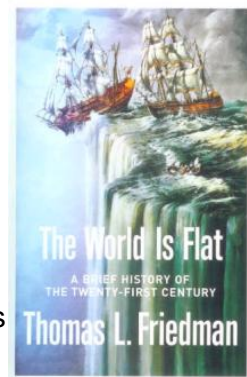
While GRACE used automated inter- and intra-GDC metadata-permitted program movement, staffers monitored system health through green-yellow-red dashboard alerts. They graphically shifted workload within and between blades for scheduled repairs and to add or remove servers and racks as needed. 1-800-FOR-GRACE customer trouble tickets and those generated by advanced and predictive analytics appeared on giant command center displays. The health of

hardware, operating software, and programs were logically compared to historical “performance norms”, allowing workers to focus on issues rather than examining “miles” of log files. For example, a performance alert might appear on the dashboard as a router malfunction rather than a bad blade. A simplified dashboard also alerted customers to severe issues.

The GRACE senior management also used the dashboard to manage their GRACE franchise investment by tracking physical assets, utilization rates, labor costs, environmental expenses, pricing models, customers’ response times, and many more business metrics.

Conclusion - Science Fiction Is a Prelude To Science Fact

If history tells us anything, it is that facts change and the ways of the past don’t stay that way forever. We either “...embrace change or risk getting left behind.”¹¹⁴ We will have to think differently about computing. After all, it took Columbus until 1492 to prove the world wasn’t flat. In 1930, Clyde W. Tombaugh proclaimed Pluto as the ninth planet, yet 76 years later we were back to eight. Throughout computer science history, computing was performed in a certain way, occupying fast data centers, restrained by issues of security and complexity. With the advent of cloud computing, secured and simplified by advances attributed to GRACE, the world awakened to a new dawn of technology.






Can science-fiction help chart the path to the future? Everything we do as a society involves planning, from deciding what to eat for breakfast to building a bridge. In the computing world, tomorrow’s ideas were envisioned years before by people who dreamt of how technology *could* work. Depending on the industry, it can take 5-10 years to go from a vision to a product, such as in the case of an automobile¹¹⁵. Designers need a vision today for the cloud. GRACE may be science fiction, but for future casting, it is clearly based on today’s technology and offers a solution to many of our perplexing issues.

We witnessed how Apple’s tablet computer dream progressed from their 1993 Newton PDA, to the iPhone in 2007, and finally the breakthrough 2010 iPad. Did Apple’s vision develop in a vacuum? Looking back, history shows that some science fiction predictions came true¹¹⁶, such as the tablet from Star Trek and Arthur C. Clarke’s 1968 “2001: A Space Odyssey” “*When he tired of official reports and memoranda and minutes, he would plug in his foolscap-size newspad into the ship’s information circuit and scan the latest reports from Earth. One by one he would conjure up the world’s major electronic*

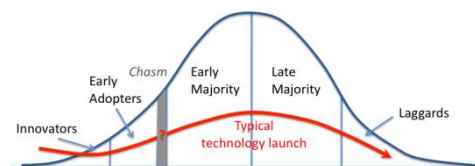


papers...”. “...when he punched that, the postage-stamp-size rectangle would expand until it neatly filled the screen and he could read it with comfort. When he had finished, he would flash back to the complete page and select a new subject for detailed examination...”.

Other examples of science-fiction that came true included:

- Earbud headphones – “Fahrenheit 451” by Ray Bradbury (1953). “*And in her ears the little seashells, the thimble radios tamped tight, and an electronic ocean of sound, of music and talk and music and talk coming in...*” 
- Atom bomb – “The World Set Free” by H. G. Wells (1914). “*Those used by the Allies were lumps of pure Carolinum, painted on the outside with unoxidised cydonator inducive enclosed hermetically in a case of membranum. A little celluloid stud between the handles by which the bomb was lifted was arranged so as to be easily torn off and admit air to the inducive, which at once became active and set up radio-activity in the outer layer of the Carolinum sphere. This liberated fresh inducive, and so in a few minutes the whole bomb was a blazing continual explosion.*”¹¹⁷ In 1934, physicist Leo Szilard patented the use of neutrons instead of fictional Carolinum in a chain reaction to split atoms as prelude to his work on the Manhattan Project’s atom bomb¹¹⁸. 
- Spaceships – “From the Earth to the Moon” by Jules Verne (1865) had similarities to the Apollo moon missions. Verne’s capsule transported three people and Apollo carried three astronauts. His launch was from central Florida, not far from Cape Kennedy, and both capsules were recovered at sea. 

As successful as GRACE turned out, it did not mark the end of traditional hardware and software. In 1999, Geoffrey Moore’s “Crossing the Chasm”¹¹⁹ illustrates that



some users are left behind and never cross the chasm, while the majority join, adopt, and embrace the technology. A small percentage of users and products did not leverage GRACE. This included fuddy-duddies who are new technology-averse, the paranoid who are probably still generating their own electricity, and products that could not be virtualized. Examples of non-virtualized technology that could not cross the chasm included real-time control nuclear reactors and processors in automobiles that helped steer the car. Others gleefully experienced dependable ease-of-use utility computing with its’ lower costs and the reduced complexity of private, public, and hybrid multi-tenant applications.

As GRACE went on, it ceased to be a platform for the migration of older applications, but the place where the next generation of applications were found. GRACE was a roadmap to facilitate the development of next-generation global initiative, the “personal cloud”, or as it was known, GRACE PC or PC for short. The PC was an extension of the private-public-hybrid models and focused on the needs of the individual. In the personal cloud, you could track your athletic performance through a chip implanted in your favorite running shoes¹²⁰. It also helped mankind through a big data analysis of health monitoring sensors, allowed car traffic to flow, helped airplanes fly safely, automatically prepared grocery lists, and in general, truly allowed information to be at their fingertips. This photo shows an early wearable sensor layered on a temporary tattoo.¹²¹



Over the next twenty years – 2040 – the power of the devices used to access GRACE exceeded the compute power of GRACE. The world turned into a giant grid and one-by-one, GRACE’s role was reduced from a place where giant compute engines were found to a network that enabled worldwide collaboration, similar to SETI@home.¹²² Global parallel processing was introduced that tied user’s devices and the global GRACE network together to solve some of the problems facing mankind such as disease and hunger.

People would come to own devices with integrated technology that communicated with and through GRACE. Based on a person’s schedule, GRACE PC suggested the appropriate clothes to wear for warmth and appearance. Purchasing tickets to a sporting event and GRACE would suggest a time to leave for the event and the best way to get there. Even crops benefited from GRACE, which could suggest the right mixture of soil and planting mix based on drainage, amount of sunshine and orientation, and the expected temperature ranges for the area. The data explosion continued and as a result, GRACE took on aspects of autonomic computing – “self-managing characteristics of distributed computing resources, adapting to unpredictable changes whilst hiding intrinsic complexity to operators and users.”¹²³ This allowed GRACE to adopt and add new functionality on its own as it continued on its journey towards a true open cloud. Grace Hopper would have been proud of how the computing world turned out.

Sub-Commander T'Pol: I still don't believe in time travel.¹²⁴

Captain Jonathan Archer: The hell you don't.



Footnotes

- ¹ "Star Trek: Enterprise: Shockwave: Part 1 (#1.26)" (2002) Vulcan First Officer/Science Officer Sub-Commander T'Pol served aboard the star ship Enterprise NX-01 in the year 2151
- ² "Science Fiction Prototyping – Designing the Future with Science Fiction" ISBN 9781608456550
- ³ "Programming Windows Azure: Programming the Microsoft Cloud" by Sriram Krishnan, p 24, ISBN 978-0596801977
- ⁴ <http://idcdocserv.com/1142>
- ⁵ http://explainingcomputers.com/cloud/BGT_Cloud_Computing_Extract.pdf
- ⁶ Transcript of embedded video, <http://www.fool.com/fool/free-report/15/rbsoundecap-67194.aspx>
- ⁷ http://en.wikipedia.org/wiki/Nikola_Tesla
- ⁸ Transcript of embedded video, <http://www.fool.com/fool/free-report/15/rbsoundecap-67194.aspx>
- ⁹ http://www.nj.com/business/index.ssf/2011/09/smartphones_overtake_feature_p.html
- ¹⁰ http://www.hp.com/hpinfo/newsroom/press_kits/2011/HPDiscover2011/DISCOVER_5_Myths_of_Cloud_Computing.pdf
- ¹¹ <http://www.kurzweilai.net/memorandum-for-members-and-affiliates-of-the-intergalactic-computer-network>
- ¹² <http://www.google.com/press/podium/ses2006.html>
- ¹³ <http://aws.amazon.com/vpc/>
- ¹⁴ http://www.rackspace.com/PPC/private_cloud.php?CMP=PPC_HybridBU_Google_private+cloud_exact
- ¹⁵ <http://blogs.vmware.com/rethinkit/2011/06/future-of-cloud-and-nyse-uronext-community-platform.html>
- ¹⁶ http://www.accountingweb-cgi.com/pastnews/wbb_030608.html
- ¹⁷ <http://www.nytimes.com/2011/06/18/technology/18security.html>
- ¹⁸ <http://i.techweb.com/darkreading/sophos/sophos-security-threat-report-2011-wpna.pdf>
- ¹⁹ <http://www.thestateofcloudcomputing.com/>
- ²⁰ <http://www.slideshare.net/wigos/facebook-present-future-by-nick-gonzalez>
- ²¹ <http://www.checkfacebook.com/>
- ²² <http://techcraver.com/2011/01/09/skype-announcements-from-ces-2011/>
- ²³ <http://www.fichier-pdf.fr/2011/09/01/vnx-storage-efficiencies-what-why-and-when/vnx-storage-efficiencies-what-why-and-when.pdf>
- ²⁴ www.cloud-council.org
- ²⁵ www.opendatacenteralliance.org
- ²⁶ CloudSecurityAlliance.org
- ²⁷ cloudforum.org
- ²⁸ opencloudmanifesto.org
- ²⁹ occi-wg.org
- ³⁰ <http://www.nytimes.com/2011/06/18/technology/18security.html?pagewanted=all>
- ³¹ <http://www.uscis.gov/portal/site/uscis/>
- ³² <http://www.cstl.nist.gov/strbase/fbicore.htm>
- ³³ <http://www.montana.edu/wwwmb/coursehome/mb105/Lectures/Chapter%208.ppt>
- ³⁴ <http://www.forbes.com/forbes/2011/0117/features-jonathan-rothberg-medicine-tech-gene-machine.html>
- ³⁵ <http://www.vmware.com/virtualization/virtual-machine.html>
- ³⁶ <http://www.crews.org/curriculum/ex/compsci/articles/generations.htm>
- ³⁷ http://en.wikipedia.org/wiki/Transistor_count
- ³⁸ <http://usmantariq.org/blog/homepage/?cat=30>
- ³⁹ http://download.intel.com/newsroom/kits/22nm/pdfs/22nm_Fun_Facts.pdf
- ⁴⁰ http://newsroom.intel.com/community/intel_newsroom/blog/2011/06/20/intel-equipped-to-lead-industry-to-era-of-exascale-computing
- ⁴¹ http://download.intel.com/pressroom/pdf/kkuhn/Kuhn_IWCE_invited_text.pdf
- ⁴² <http://www.cpu-wars.com/2010/10/amd-says-that-cpu-core-race-cant-last.html>
- ⁴³ <http://newsroom.intel.com/docs/DOC-2032>
- ⁴⁴ <http://computer.howstuffworks.com/small-cpu1.htm>
- ⁴⁵ <http://electronicsistechology.blogspot.com/2011/01/tri-gate-transistor.html>
- ⁴⁶ http://download.intel.com/newsroom/kits/22nm/pdfs/22nm-Details_Presentation.pdf
- ⁴⁷ http://www.informationweek.com/news/hardware/processors/232200660?cid=RSSfeed_IWK_News
- ⁴⁸ <http://newscenter.berkeley.edu/2011/09/12/ferroelectrics-used-for-negative-capacitance/>
- ⁴⁹ <http://www.x-drivers.com/news/hardware/4338.html>
- ⁵⁰ http://www.intechopen.com/source/pdfs/16177/InTech-Future_memory_technology_and_ferroelectric_memory_as_an_ultimate_memory_solution.pdf
- ⁵¹ <http://dvice.com/archives/2010/03/3d-chips-to-tak.php>
- ⁵² http://news.cnet.com/8301-17938_105-20066890-1/nasa-to-demonstrate-super-cool-cooling-technology/
- ⁵³ <http://www.amd.com/us/products/server/processors/6000-series-platform/6200/Pages/6200-series-processors.aspx>
- ⁵⁴ "The Future of Computing Performance: Game Over or Next Level?" by Samuel H. Fuller and Lynette I. Millett, ISBN 978-0-309-15951-7, page 91
- ⁵⁵ <http://techresearch.intel.com/ProjectDetails.aspx?Id=151>
- ⁵⁶ <http://www.magneticdiskheritagecenter.org/100th/Progress/Hoagland/SLIDE11.PDF>

57 http://www.bitsavers.org/pdf/ibm/38xx/3830/GA26-1592-2_3830_3330_Ref_Apr72.pdf
58 http://www.seagate.com/www/en-us/about/corporate_information/company_milestones
59 <http://www.seagate.com/support/disc/manuals/scsi/29170b.pdf>
60 <http://www.seagate.com/support/disc/manuals/scsi/83329484e.pdf>
61 http://en.wikipedia.org/wiki/Heat-assisted_magnetic_recording
62 <http://www.violin-memory.com/assets/Violin-WP-ServerStorage-Performance-Gap.pdf?d=1>
63 <http://www.violin-memory.com/assets/Violin-WP-Disk-Storage-Shortfall.pdf?d=1>
64 This simplification does not account for drive optimization techniques such as command queuing. Command queuing prioritizes I/O requests and services them based on current disk head location on the rotating disk platter, thereby reducing avg. service time.
65 <http://en.wikipedia.org/wiki/IOPS>
66 http://www.ocztechnology.com/images/products/downloads/SSDvsHDD_quick_reference_9.pdf
67 http://www.storagereview.com/origin_solid_state_drives
68 http://en.wikipedia.org/wiki/EMC_Symmetrix#Fully_Automated_Storage_Tiering_.28FAST.29
69 <http://www.seagate.com/ww/v/index.jsp?vgnextoid=afb2308aaecb8210VgnVCM1000001a48090aRCRD>
70 www2.engr.arizona.edu/~ece369/SSD.ppt
71 http://www.supertalent.com/datasheets/SLC_vs_MLC%20whitepaper.pdf
72 http://www.flashmemorysummit.com/English/Conference/TwentyYears_Flash.html
73 http://www.theregister.co.uk/2011/11/09/emc_lightning_strike/
74 http://en.wikipedia.org/wiki/Phase-change_memory
75 http://www.computerworld.com/s/article/9218031/IBM_announces_computer_memory_breakthrough
76 <http://www.originami.com/sp/milestones.htm>
77 <http://cloudo.com/>
78 <http://www.jolicloud.com/>
79 <http://eyeos.org/>
80 <http://amazonsilk.wordpress.com/2011/09/28/introducing-amazon-silk/>
81 http://en.wikipedia.org/wiki/History_of_the_iPhone
82 <http://www.investmentu.com/2011/June/icloud-and-cloud-computing.html>
83 http://en.wikipedia.org/wiki/Video_on_demand#History
84 <http://www.tnr.com/article/the-pc-officially-died-today>
85 http://www.dnv.com.br/binaries/TO2020%20lowres%20SMALLERf_tcm156-453252.pdf
86 <http://mc10inc.com/>
87 <http://query.nytimes.com/mem/archive/pdf?res=F30815FE3B58117B93CBA9178AD85F438685F9>
88 <http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf>
89 <http://www.emc.com/collateral/emc-perspective/h2159-managing-storage-ep.pdf>
90 Joseph Tucci, Chairman, President and CEO EMC Corporation, Keynote Address, EMC World 2011
91 http://library.ltu.edu.tw/download/periodical/20100419_InformationWeek.pdf
92 http://www.businessweek.com/magazine/content/06_05/b3969401.htm
93 <http://www.catalinamarketing.com/home.php>
94 http://www.aristanetworks.com/media/system/pdf/SwitchingArchitecture_wp.pdf
95 http://www.computerworld.com/s/article/358854/Full_Duplex_Boosts_Network_Traffic
96 <http://topics.dallasnews.com/article/06ft5M80MdbvG?q=First+Data>
97 http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns1175/Cloud_Index_White_Paper.html
98 http://en.wikipedia.org/wiki/History_of_mobile_phones
99 http://www.privateline.com/mt_cellbasics/2006/01/cell_and_sectorterminology.html
100 http://www.ctia.org/consumer_info/index.cfm/AID/10323
101 <http://www.chetansharma.com/blog/2011/08/15/ny-times-as-networks-speed-up-data-hits-a-wall/>
102 <http://cims.clayton.edu/sakhtar/EncyclopediaPaper.pdf>
103 <http://www.bbc.co.uk/news/technology-15467740>
104 <http://en.wikipedia.org/wiki/Photophone>
105 <http://bits.blogs.nytimes.com/2011/07/18/using-light-to-send-data-across-the-room/>
106 <http://www.nytimes.com/2011/06/10/technology/10chip.html>
107 "Telecommunications and Data Communications Handbook" by Ray Horak, p470. ISBN 978-0-470-04141-3"
108 <http://www.wired.com/wiredenterprise/2011/12/nonexistent-supercomputer/all/1>
109 http://sas70.com/sas70_faqs.html
110 <http://aws.amazon.com/ec2-sla/>
111 <http://www.intel.com/content/www/us/en/cloud-computing/cloud-computing-cloud-vision-2015-video.html.html>
112 <http://www.fujitsu.com/global/news/pr/archives/month/2011/20111109-04.html>
113 <http://www.aristanetworks.com/media/system/pdf/CloudNetworkLatency.pdf>
114 http://www.thebottomlinenews.ca/documents/TBL_FOCUS_Education%20for%20Accountants.pdf
115 <http://www.autofieldguide.com/articles/scion-iq-sized-up>
116 <http://mashable.com/2010/09/25/11-astounding-predictions/>
117 "The World Set Free" by H. G. Wells, 1914, pps 100-101, First Edition.
118 "H.G. Wells" by Harold Bloom, ISBN 0-7910-8130-3, Page 5
119 Publisher: Harper Paperbacks, ISBN 978-0060517120
120 <http://www.informationweek.com/news/global-cio/interviews/232200563>
121 <http://www.nytimes.com/2011/09/04/technology/wireless-medical-monitoring-might-untether-patients.html>

¹²² "With over 278,832 active computers in the system (2.4 million total) in 234 countries, as of November 14, 2009, SETI@home has the ability to compute over 769 teraFLOPS. For comparison, the K computer, which as of June 20th 2011 was the world's fastest supercomputer, achieved 8162 teraFLOPS." Source: <http://en.wikipedia.org/wiki/SETI@home>

¹²³ http://en.wikipedia.org/wiki/Autonomic_Computing

¹²⁴ "Star Trek: Enterprise: Shockwave: Part 2 (#2.1)" (2002)

EMC believes the information in this publication is accurate as of its publication date. The information is subject to change without notice.

THE INFORMATION IN THIS PUBLICATION IS PROVIDED "AS IS." EMC CORPORATION MAKES NO REPRESENTATIONS OR WARRANTIES OF ANY KIND WITH RESPECT TO THE INFORMATION IN THIS PUBLICATION, AND SPECIFICALLY DISCLAIMS IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.

Use, copying, and distribution of any EMC software described in this publication requires an applicable software license.