

Future of Computing is Boring*

Aleksander Slominski
Vinod Muthusamy
Vatche Ishakian

IBM Research

***and that is
exciting!**

Disclaimer

“It's tough to make predictions, especially about the future.”

- Yogi Berra

Alek Corollary: “My predictions are most certainly wrong.”

- Alek

Quick History of Computing

Citroen C3 5CV

— — —
France 1922

Engine: 4-Cylinder

Displacement: 856 cc

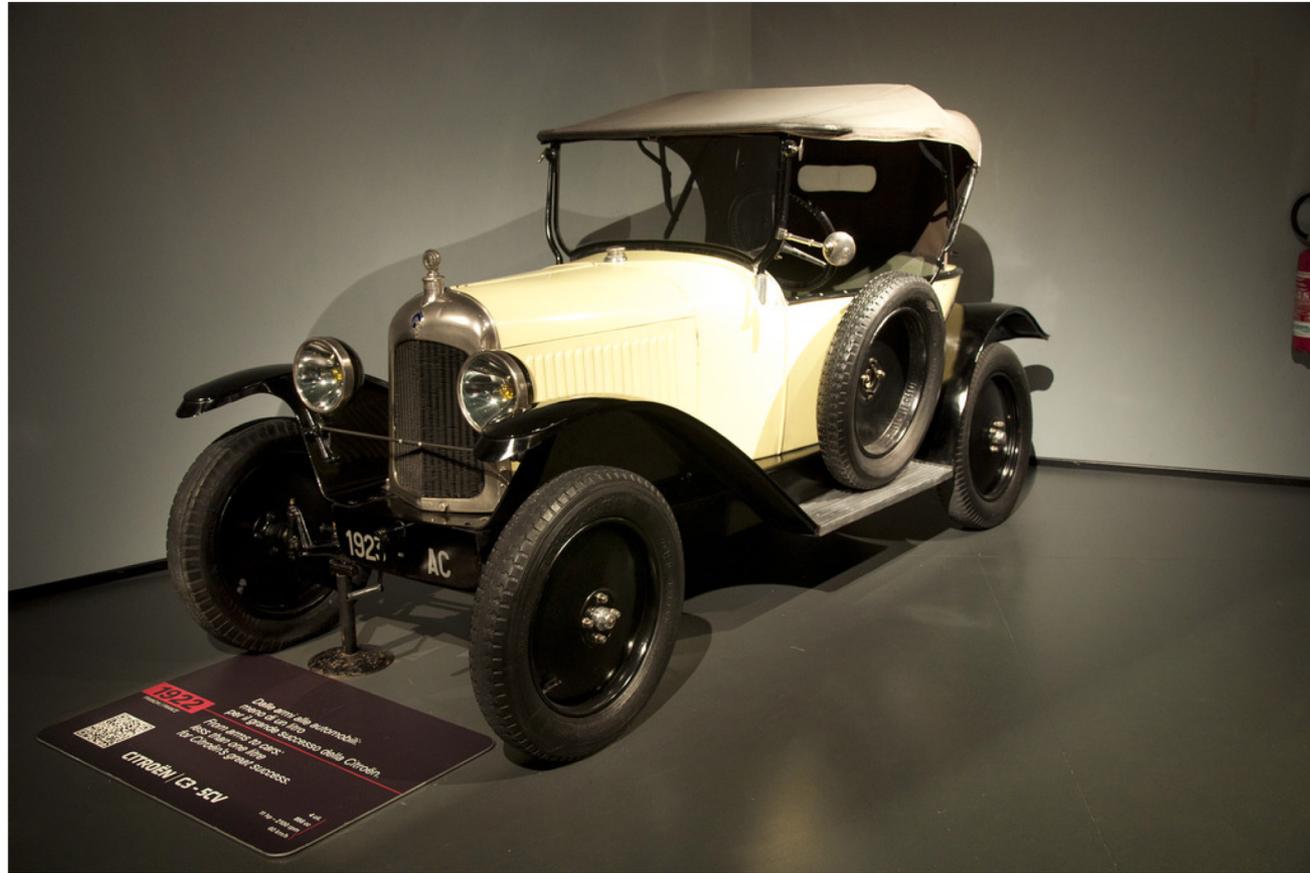
Max Power: 11 HP @ 2100 rpm

Speed: 60 km/h

Weight: 420 kg

Cost around: \$20K

(for today Citroen C3 2017 £11K-£17K)



<https://www.flickr.com/photos/132623682@N03/18026248601>

What if cars improved like computers? In 50 years

1922

Max Power: 11 HP @ 2100 rpm

Speed: 60 km/h

Weight: 420 kg

Cost: \$20K

1972

Max Power: ???

Speed: ???

Weight: ???

Cost: ???

Cost of computation changes in 50 years

— — —

Year	Approximate cost per GFLOPS (2017 US dollars)	Approximate cost to buy 1 GB of memory
1960	\$150B*	\$5B*
1980	\$40M	\$6M
2000	\$1K	\$840
2010	\$2**	\$20
2020 (extrapolated from 2018)	\$0.02**	\$7

* This is an extrapolated cost as there were hardware limitation on getting any machine with that level of performance. (\$1B = 10^9 US dollars)

** Main advance in computation are by combining CPU with GPUs.

“Cost of computing - Approximate cost per GFLOP” https://en.wikipedia.org/wiki/FLOPS#Hardware_costs

”Memory Prices (1957-2018)” <https://jcmmit.net/memoryprice.htm>

Computing for Moon Landing (1960)

IBM 7090

In 1960, a typical system sold for \$2.9 million (**equivalent to \$18 million in 2017**) or could be rented for **\$63,500 a month (equivalent to \$405,000 in 2016)**.

https://en.wikipedia.org/wiki/IBM_7090

IBM 7090 speed was described as “perform[ing] any of the following operations in one second: 229,000 additions or subtractions, 39,500 multiplications, or 32,700 divisions.”

“Hidden Figures” Computers

Jean Bartik, one of the earliest pioneering women in technology, talks about her memories of breaking into the then new field of computer science and working on the ENIAC in the 1940s.”

<https://www.computerhistory.org/atcm/hidden-figures-no-longer/>

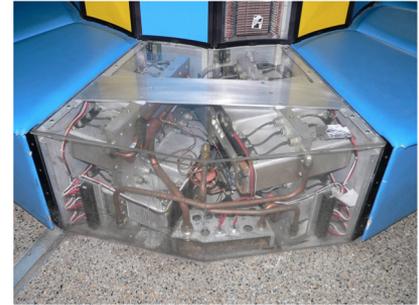


Cray X-MP (1980)

1984 X-MP/48 costs about US\$15 million plus the cost of disks. In 1985 Bell Labs purchased a Cray X-MP/24 for **\$10.5 million** along with **eight DD-49 1.2 GB drives** for an **additional \$1 million**.

Each CPU had a theoretical peak performance of 200 MFLOPS, for a **peak system performance of 400 MFLOPS**.

https://en.wikipedia.org/wiki/Cray_X-MP



Bunyip Beowulf cluster (2000)

“Beowulf style Linux cluster called Bunyip at the Australian National University, Canberra. This cluster is built with 98 dual 550 MHz Pentium III nodes, each equipped with 384 Megabytes of RAM (total about 36 Gigabytes), 13 Gigabytes of disk space (total 1.3 Terabytes) and 3 × 100 MBit/s fast Ethernet cards”

“Parallel Data Mining on a Beowulf Cluster” (2001)

Cost about $100 \times 600 = \$60K$ (“\$600 for each basic tower node”)

<https://www.wired.com/2000/12/beowulf/>



The Borg, a 52-node Beowulf cluster used by the McGill University pulsar group to search for pulsations from binary pulsars

https://en.wikipedia.org/wiki/Beowulf_cluster

Sony PlayStation 4 (2013)

Sony PlayStation 4 is listed as having a peak performance of **1.84 TFLOPS**, at a price of **\$400**



The PlayStation 4 uses an Accelerated Processing Unit (APU) developed by AMD in cooperation with Sony. It combines a central processing unit (CPU) and graphics processing unit (GPU), as well as other components such as a memory controller and video decoder. The CPU consists of two quad-core Jaguar modules totaling 8 x86-64 cores, 7 of which are available for game developers to use. The GPU consists of 18 compute units to produce a theoretical peak performance of 1.84 TFLOPS. The system's GDDR5 memory is capable of running at a maximum clock frequency of 2.75 GHz (5500 MT/s) and has a maximum memory bandwidth of 176 GB/s. The console contains 8 GB of GDDR5 memory

https://en.wikipedia.org/wiki/PlayStation_4

50 Teraflops Box (2017)

Part List:

- 2-AMD Radeon Vega Frontier Edition 16GB GPU \$2,000
 - 1-G.SKILL 32GB (2 x 16GB) DDR4 Memory \$200.99
 - 1-Crucial 525GB SATA III 3-D SSD \$159.99
 - 1-EVGA Plus 1000W Gold Power Supply \$119.99
 - 1-MSI X370 AM4 AMD Motherboard \$129.99
 - 1-AMD RYZEN 7 1700 8-Core 3.0 GHz (3.7 GHz Turbo) Processor \$3
 - 1-TOSHIBA 5TB 7200 RPM SATA Hard Drive \$146.99
 - 1-Rosewill ATX Mid Tower Case \$49
- Total: \$3,122 or a mind-boggling \$62 per teraflop!**

The box that we are assembling will be **capable of 50 teraflops (fp16 or half precision) ... for \$3K**

Building a 50 Teraflops AMD Vega Deep Learning Box for Under \$3K

<https://medium.com/intuitionmachine/building-a-50-teraflops-amd-vega-deep-learning-box-for-under-3k-ebdd60d4a93c>



How much improved?

CPU = $3750 * 40000 * 500 * 100 = 7.5e+12 \approx 10^{12} = 10T$
(M,B,T)

= 10 Billions in Long scale (Western, Central Europe, older British, and French Canadian)

= 10 Trillions in Short scale (US, Eastern Europe, English Canadian, Australian, and modern British)

https://en.wikipedia.org/wiki/Names_of_large_numbers

Memory $800 * 7437 * 42 * 3 = 74\,9649\,600 \approx 100M$ improvement

Changes in Cost and Weight

IBM 7090 (1959)

Weight: 70,000 pounds (35 short tons; 32 t)

Cost: \$18M https://en.wikipedia.org/wiki/IBM_7030_Stretch

Teraflops Box (2017)

Weight (with monitor): ~ 30 pounds (13.6Kg) (ratio 70K/30 = 70000/30 = 2333x)

Cost: \$3K (ratio 18M/3K = 18000/3 = 6000x)

What if cars improved like computers? In 50 years you get:

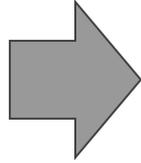
1922

Max Power: 11 HP @ 2100 rpm

Speed: 60 km/h

Weight: 420 kg

Cost: \$20K



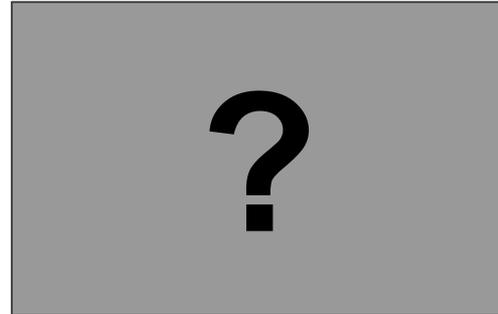
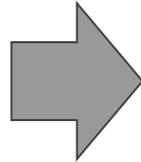
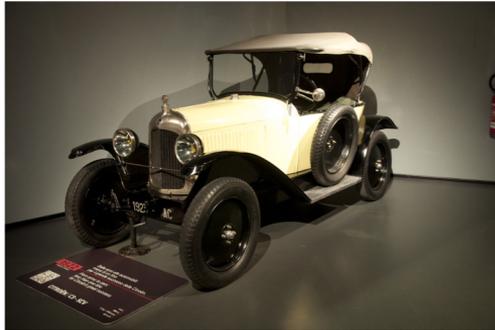
1972

Max Power: ???

Speed: ???

Weight: ???

Cost: ???



What if cars improved like computers? In 50 years you get:

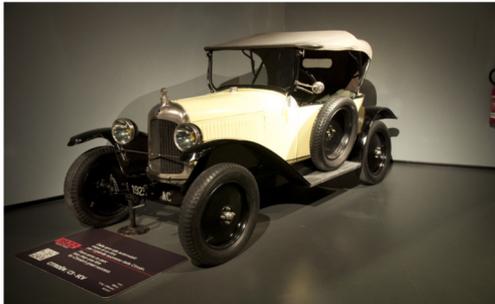
1922

Max Power: 11 HP @ 2100 rpm

Speed: 60 km/h

Weight: 420 kg

Cost: \$20K



1972

Max Power: 11 Trillions (or Billions) HP

Speed: 60 Trillions (or Billions) km/h

Speed of light: 1 079 252 848 km/h =

~ 1B (or 1 Milliard) km /h

Weight: 0.18kg = 0.4 pounds =

6 oz



Cost: \$3.3

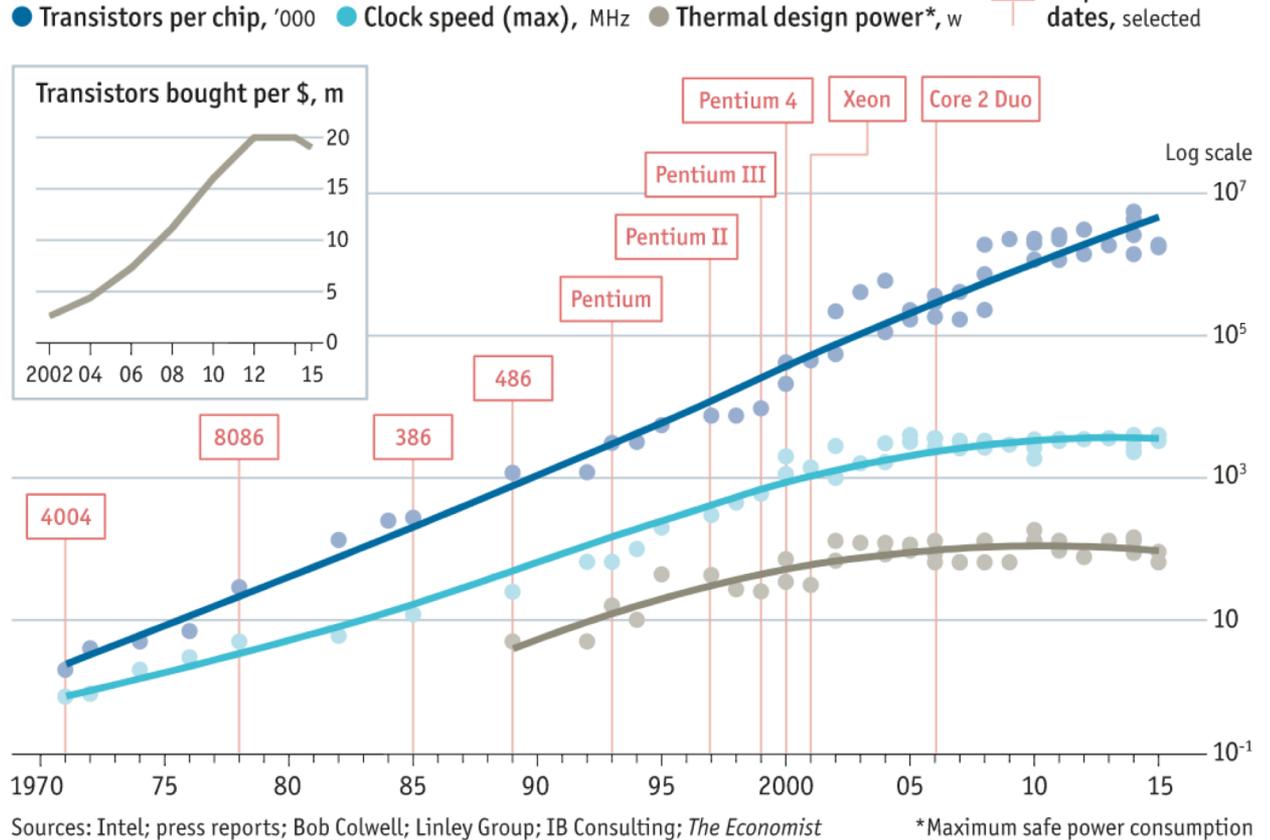


Moore's Law

— — —

“After a glorious 50 years, Moore’s law—which states that computer power doubles every two years at the same cost—is running out of steam.”

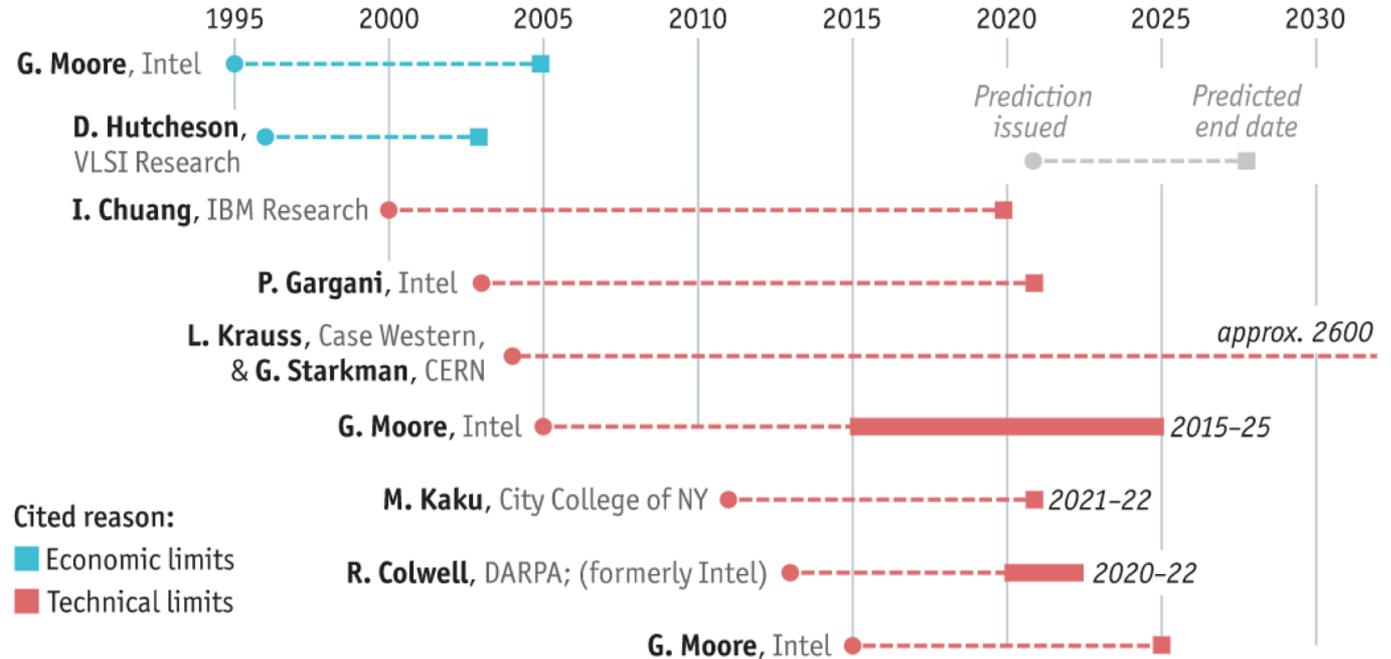
Stuttering



End?

Faith no Moore

Selected predictions for the end of Moore's law



Sources: Intel; press reports; *The Economist*

Computing vs Humans

Cost of Original Computers aka Developers

- There is about 1500 to 2000 hours of business hours in one year
 - 50 weeks x 40 hours more or less depending on vacations, holidays, etc.)
- The cost per hour?
 - from \$20 to \$100 on average
 - → **the cost of developer per year to be around \$40K to \$200K/year**
 - in western countries, world-wide may be lower ...
- Back-of-envelope calculations show that the cost is high?

Number of developers worldwide

- Demand for programmers keep growing outpacing supply
- Estimated there is around **20 millions of professional developers worldwide in 2017**
 - “Software Developer Statistics: How Many Software Engineers Are There in the US and the World?”, Oct 31, 2017, <https://www.daxx.com/article/software-developer-statistics-2017-programmers>

Changes in cost of developers

- Considering college graduate average salary over last 60 years it changed from \$6K in 1960 to \$50K in 2015 - that is 8x increase
- **Adjusted for inflation the increase is only 1.06x (that is 6% increase)**
 - “Salary Trends Through Salary Survey: A Historical Perspective on Starting Salaries for New College Graduates”, August 02, 2016 <http://www.naceweb.org/job-market/compensation/salary-trends-through-salary-survey-a-historical-perspective-on-starting-salaries-for-new-college-graduates/>
- Average cost of one developer → \$100K / year

How much computation for \$100K?

Year	How many GFLOPs for \$100K in 2017 US dollars?	How many GBs for \$100K?
1960	0.0000006 GFLOPS or 600 FLOPS	0.00002 GB or 20KB
1980	0.0025 GFLOPS (3750x)	0.016 GB (800x)
2000	100 GFLOPS (40000x)	119 GB (7437x)
2010	50000 GFLOPS or 50 TFLOPS (500x)	5000 GB or 5 TB (42x)
2020 (extrapolated from 2018)	5M GFLOPS or 5 PFLOPS (100x)	15000 GB or 15 TB (3x)

Cost of Computation and Cost of Developers

- Cost of computation: **several orders of magnitude more** computational resources can be bought with the same amount of money in 10-20 years ...
- Cost of developers: human resources
 - if anything time is getting more expensive when compared to how much compute time can be bought ...

Future of Computing (Learning from Past)

Evolution of computation

Year	Computing Paradigm	Expenses	Time to solution	Cost of compute with 1 GB of memory per hour
1960-2000	Mainframes, data centers	Upfront capital investment	Months to Years	\$100K to \$1000
2000s	HPC Cluster and Grid Computing, data centers, VMs	Capital investment, reusing idle machines	Days to Weeks	\$10-50
2010s	Cloud and Serverless Computing, VMs and Containers	Pay-as-you-go, sometimes paying even if not using compute resources (VMs)	Minutes to Hours	\$0.01 - 0.06
Future	Standardized containers?	Pay only for what is used	Seconds to Minutes	<\$0.01

“Comparing Cloud Instance Pricing: AWS vs Azure vs Google vs IBM”, November 18, 2017, <https://www.rightscale.com/blog/cloud-cost-analysis/comparing-cloud-instance-pricing-aws-vs-azure-vs-google-vs-ibm>

Rise of Containers

- Open Virtualization Format
<https://www.dmtf.org/standards/ovf>
- The future of Linux Containers, **2013**
<https://www.youtube.com/watch?v=wW9CAH9nSLs>
- Open Container Initiative (OCI)
<https://www.opencontainers.org/>
- Firecracker - Lightweight Virtualization for Serverless Computing, **2018**
<https://github.com/firecracker-microvm/>
- Future ...?

Real-world Shipping Containers

Success story

(Short) History of shipping containers - success story

“In the early 1950’s McLean began his own “container concept”

Matson, on the West coast of the US also attempted a type of container concept, but failed sorely.

In the late 1960’s Sea-Land got some large boosts from the US Government and US Military.

McLean and Sea-Land were able to finally standardize the container concept in the early 70's.

“The History Of ISO Shipping Containers From Wooden Crates To Steel Boxes”

<http://www.isbu-association.org/history-of-shipping-containers.htm>

(Short) History of shipping containers - success story

In 1956, loose cargo cost **\$5.86 per ton to load**. Using an ISO shipping container, the cargo cost was **reduced to only 16 cents per ton**.

$$5.86 / 0.16 = 36x$$

There were many who had similar concepts previously but McLean was simply the guy, who with the push of the US military, really made the "standardized container" concept spread globally.

“The History Of ISO Shipping Containers From Wooden Crates To Steel Boxes”
<http://www.isbu-association.org/history-of-shipping-containers.htm>

(Short) History of shipping containers - success story

For the first few years, from 1956 to 1965 virtually all use of the shipping containers were only by McLean on his own ships.

We must remember, the ISO shipping container as we know it today was developed in 1956, but was not re-designed and standardized Internationally until in 1972.

The standardization process which began by the ISO and IMO in 1967, was only pushed-forward at the urging of the US Government for their use by the US military for transport and housing overseas.

“The History Of ISO Shipping Containers From Wooden Crates To Steel Boxes”

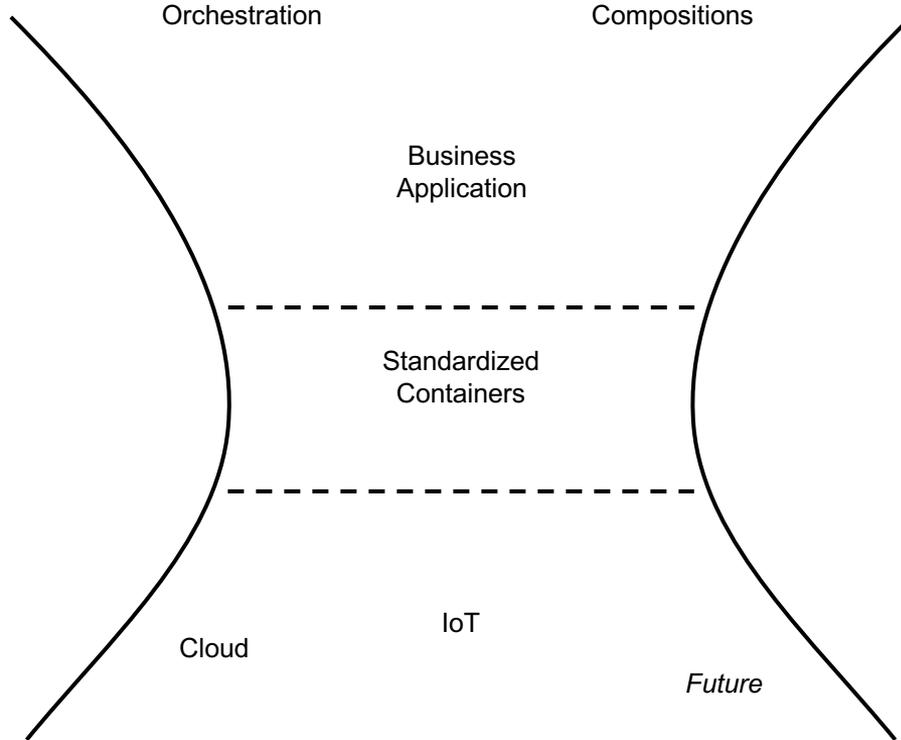
<http://www.isbu-association.org/history-of-shipping-containers.htm>

Inevitability of progress?

“People are mistaken when they think that technology just automatically improves. It does not automatically improve. **It only improves if a lot of people work very hard to make it better and I actually, I think, by itself, degrade, actually.** You look at ancient civilizations like ancient Egypt and they were able to make the pyramids and they forgot how to do that. And the Romans, they built these incredible aqueducts, they forgot how to do it.

- Elon Musk

Universal Standardized Computing Container like Electricity?



STANDARDS



< PREV

RANDOM

NEXT >



HOW STANDARDS PROLIFERATE:

(SEE: A/C CHARGERS, CHARACTER ENCODINGS, INSTANT MESSAGING, ETC.)



Future of Computing is about ...

Summary: Time is money (use it wisely)

There is about 1500 to 2000 hours of work hours in one year

- 50 weeks x 40 hours (more or less depending on vacations, holidays, etc)

The cost of a developer ranges from \$20 to \$100 per hour on average resulting in the cost of a developer per year to be around \$40K to \$200K in Western countries

How to optimize developer time?

What developers want to do?

Write code that matters?

Avoid undifferentiated heavy lifting?

(leave it to experts?)

Serverless?

What developers want to do?

Write code that matters

Avoid undifferentiated heavy lifting

(leave it to experts)



Google Cloud Functions



IBM Cloud Functions



Apache OpenWhisk



serverless.com

What one hour (\$20 and \$100) of developer time buys?

In 1 hour one developer produces N Lines of code?

40-50 lines with 1-2 bugs **per day** (15-50 bugs every 1000 lines)

\$20 buys about 10M requests each 100ms for serverless computing

$\$0.000001667 * 10M = \$16.67 + \$2$ for 10M requests

AWS Lambda costs \$0.20 per 1M requests where each request takes 100ms and gets 1GB memory/compute in serverless and cost

\$0.000001667 per 100ms with GB of memory

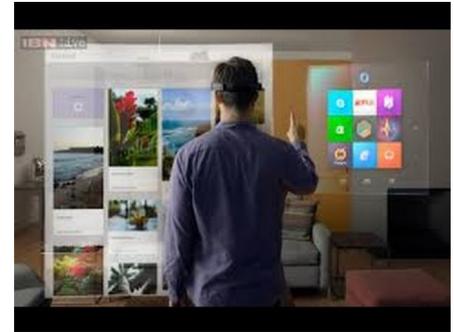
(not counting free tier)

Optimizing Cost of Computing for Productivity of Humans?

What can increase productivity?

In movies and books? Just need more Augmented or Virtual Reality (AR/VR)?

What can help bring developers together and allow collaboration (such as virtual pair programming) that otherwise would be more expensive?



Today Computing becomes “Legacy” Computing in Future

Legacy computing containerized, perhaps as virtual machines (VMs) inside containers as from an economical point of view, it may be cheaper to keep legacy code running than pay for developers to re-design and re-write code using new computing approaches.

Who builds and operates electric power plants?

Around the world, there are about 62,500 power plants operating today.

https://www.washingtonpost.com/news/wonk/wp/2012/12/08/all-of-the-worlds-power-plants-in-one-handy-map/?noredirect=on&utm_term=.af263d4f8a01

“In 2017, it was estimated that the number of data centers globally had fallen to 8.4 million.”

Predicted 7.2M in 2021 (about 500 “hyper-scale” datacenters)

<https://www.statista.com/statistics/500458/worldwide-datacenter-and-it-sites/>

Predictions (may or may not become true ...)

In the future computing works like an electric utility?

Pay-as-you-go only way to go for computation?

Like electricity today, it is just there, and we no longer think about it unless it is not working?

What will we call this future computing infrastructure?

Computing Fabric? Computing utility?

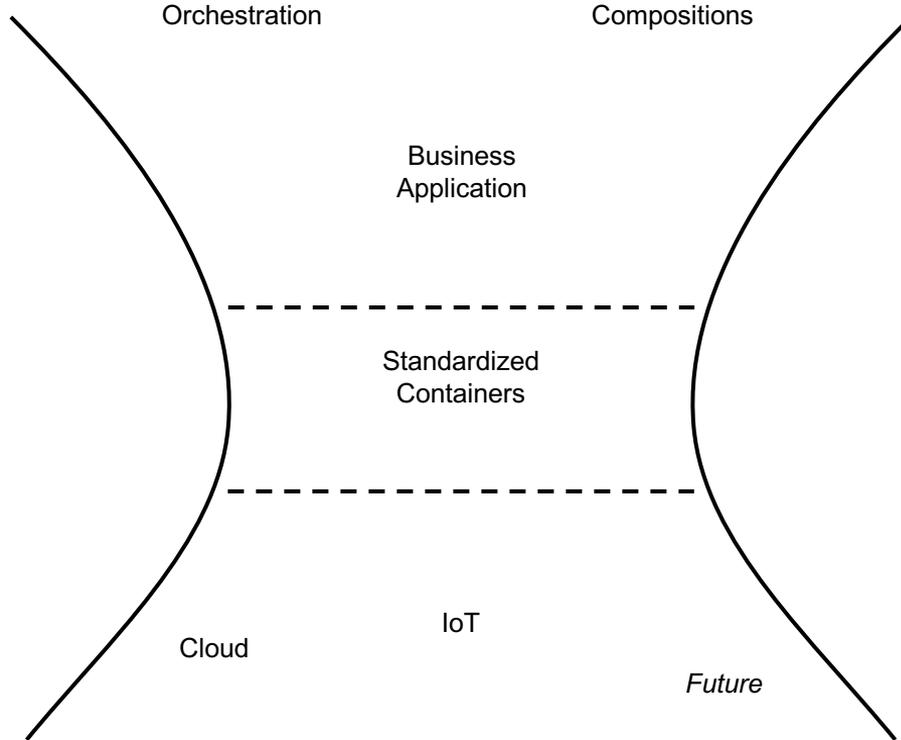
Perhaps simply computing?

What will we do with computing in future?

When computing becomes boring, the future will be exciting?!

(Boring as electricity? Invisible, available everywhere, easy to use, fast and cheap to use and to build new apps, services, ...)

Universal Standardized Computing Container like Electricity?



Questions?

Advertisement: WoSC 2019 - Fifth International Workshop on Serverless Computing (WoSC5) **CFP Submission August 30, 2019**

- Middleware, Dec 9-13, 2019 in UC Davis, CA, USA

<https://www.serverlesscomputing.org/wosc5/>

BACKUP (+30min)

Spot the differences: 1987 vs 2015

Almost 30 years difference

People in future will ...

Visions of future - tablet

Microsoft: Productivity Future Vision (2015)

<https://www.youtube.com/watch?v=w-tFdreZB94>

Productivity Future Vision (2009)

<https://www.youtube.com/watch?v=t5X2PxtvMsU>

Apple's Future Computer: The Knowledge Navigator (1987) - **+25 years ago**

<https://www.youtube.com/watch?v=hb4AzF6wEoc>

with intro and another clip

<https://www.youtube.com/watch?v=9bjve67p33E>



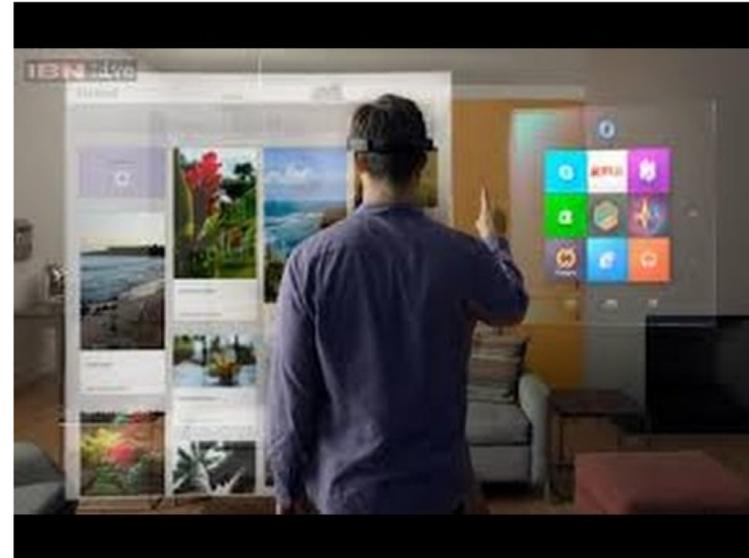
Vision of future - AR/VR

Minority Report (2002)

<https://www.youtube.com/watch?v=PJqbivkm0Ms>

Microsoft HoloLens (2015)

<https://www.youtube.com/watch?v=qym11JnFQBM>



Maneki Neko by Bruce Sterling (1998 → 2020?)

— — —

Computing that makes people happy

“ . . . Digital panarchies. Segmented, polycephalous, integrated influence networks. What about all these free goods and services you’re getting all this time?” She pointed a finger at him. “Ha! Do you ever pay taxes on those? Do you ever declare that income and those benefits? All the free shipments from other countries! The little homemade cookies, and the free pens and pencils and bumper stickers, and the used bicycles, and the helpful news about fire sales . . . You’re a tax evader! You’re living through kickbacks! And bribes! And influence peddling! And all kinds of corrupt off-the-books transactions?”

Tsuyoshi blinked. “Look, I don’t know anything about all that. I’m just living my life.”

“Well, your network gift economy is undermining the lawful, government approved, regulated economy!”

“Well,” Tsuyoshi said gently, “maybe my economy is better than your economy.”

“Says who?” she scoffed. “Why would anyone think that?”

“It’s better because we’re happier than you are. What’s wrong with acts of kindness? Everyone likes gifts. Midsummer gifts. New Years Day gifts. Year-end presents. Wedding presents. Everybody likes those.”

<http://www.lightspeedmagazine.com/fiction/maneki-neko/>

Rainbows End by Vernor Vinge (2006 → 2025)

Human / Computer Intelligence Pools

“Red doubt was hemorrhaging across the analyst pool, spreading from a statistical analysis team at Moscow-Capetown. These were the same chaps who had been consistently right about the Soybean Futures Plot. They had credibility ... and they claimed the views from the north side of the GenGen area were corrupt. Those were not views Alfred had subverted, the ones at the heart of his private operation. For better or worse, his colleagues had discovered some other deception. Now the signals and stat people in all the analyst pools had precedence. **A thousand specialists, who a second ago might have been looking at a dozen other problems, were suddenly watching the same data. Computing resources shifted from a myriad drudge tasks, began correlating data from the accessible sensors in the labs. It was as if Indo-European Intelligence were an immense cat suddenly come alert, listening and watching for sign of its prey.**

Final scene with USB stick

“Tommie reached into his jacket and pulled out a three-inch-square piece of plastic. "Here. A present for you, that **cost me all of \$19.99.**" Robert held up the dark plastic. **It looked a lot like the diskettes he'd used on his old PC at the turn of the century. He pointed a query at it. Labels floated in the air: Data Card. 128PB capacity. 97% in use.** There was more, but Robert just looked back at Tommie. "Do people still use removables like this?"

Computing in Daemon/Freedom™ (2009 → 2030?)

AR glasses, Human-centric Internet, Shamanic interface, AutoM8s Self-driving killer cars, Razorbacks, ...

“The shamanic interface is the mechanism for interacting with the darknet. **It’s called the shamanic interface because it was designed to be comprehensible to all people on earth, regardless of technological level or cultural background.**” She made a series of precise flourishes with her hands, leaving behind glowing lines in D-Space that formed an intricate pattern. As she finished, an unearthly, angelic voice sounded in the room, like a good spirit.

Sebeck looked around him for the origin of the disembodied voice.

Riley lowered her hands. “It was a hypersonic sound, Sergeant. Linked to a macro that I created based on somatic gestures. But my point is that it looks like magic. Even the most remote tribes in Papua New Guinea understand the concept of magic—and that certain rituals must be observed to invoke it. They believe in a spirit world where ancestors and supernatural beings watch over them. The shamanic interface simply connects high technology to that belief system, granting ‘powers’ and equipment as a reward for useful, organized activity.”

Taklamakan by Bruce Sterling (1999 → 2030?)

AI Evolution - Growing and Gardening instead of Design?

“Well, you could study other people's can-openers and try to improve the design. Or else you could just set up a **giant high-powered virtuality with a bunch of virtual cans inside it**. Then you make some can-opener simulations, that are basically blobs of goo. They're simulated goo, but they're also programs, and those programs trade data and evolve. Whenever they pierce a can, you reward them by making more copies of them. You're running, like, a million generations of a million different possible can-openers, all day every day, in a simulated space.”

The concept was not entirely alien to Spider Pete. "Yeah, I've heard the rumors. **It was one of those stunts like Artificial Intelligence. It might look really good on paper, but you can't ever get it to work in real life**

...

None of the creatures bothered him. He had become one of them now. **His equipment had fallen among them, been absorbed, and kicked open new doors of evolution**. Anything that could breed a can-opener could breed a rock chock and a piton, a crampon, and a pulley, and a carabiner. His haul bags, Katrinko's bags, had been stuffed with generations of focused human genius, and it was all about one concept: UP. Going up. Up and out.

BLIT by David Langford (1988 → 2000?)

Human is computer and may be hacked

“ . . . This first example of the Berryman Logical Image Technique (hence the usual acronym BLIT) evolved from AI work at the Cambridge IV supercomputer facility, now discontinued. V. Berryman and C.M. Turner [3] hypothesized that pattern-recognition programs of sufficient complexity might be vulnerable to "Gödelian shock input" in the form of data incompatible with internal representation. Berryman went further and suggested that the existence of such a potential input was a logical necessity ...

... informational analysis adopts a somewhat purist mathematical viewpoint, whereby BLITs are considered to encode Gödelian "spoilers", implicit programs which the human equipment cannot safely run. In his final paper [3] Berryman argued that although meta-logical safety devices permit the assimilation and safe recognition of self-referential loops ("This sentence is false"), **the graphic analogues of subtler "vicious circles" might evade protective verbal analysis by striking directly through the visual cortex.**

... it might be simpler to conclude that multiple simultaneous emergence of the BLIT concept was inevitable at the stage of AI research which had been reached. **The losses amongst leading theorists, in particular those with marked powers of mathematical visualization, constitute a major hindrance to further understanding ...**

Permutation City by Greg Egan

Human is computer (simulated) and may be hacked

“Never inhabit the real world again ... unless his cheapskate original scraped up the money for a telepresence robot – in which case he could spend his time blundering around in a daze, trying to make sense of the lightning-fast blur of human activity. **His model-of-a-brain ran seventeen times slower than the real thing.** Yeah, sure, if he hung around, the technology would catch up, eventually – and seventeen times faster for him than for his original. And in the meantime? He’d rot in this prison, jumping through hoops, carrying out Durham’s precious research – while the man lived in his apartment, spent his money, slept with Elizabeth...

Other visions?

Future Computation as Foundation of Future Lives?

“A reader lives a thousand lives before he dies.
The man who never reads lives only one.

-- George R. R. Martin

Reader becomes Gamer, VR-er, AR-er, ... ?

Cost of Original Computers aka Developers

The average salary of a college graduate over last 60 years changed from \$6K in 1960 to \$50K in 2015, an 8x increase.

However, if adjusted for inflation, the increase is only 6% (1.06x)

“Salary Trends Through Salary Survey: A Historical Perspective on Starting Salaries for New College Graduates”, August 02, 2016

<http://www.naceweb.org/job-market/compensation/salary-trends-through-salary-survey-a-historical-perspective-on-starting-salaries-for-new-college-graduates/>

Around 20 million professional developers worldwide in 2017

“Software Developer Statistics: How Many Software Engineers Are There in the US and the World?”, Oct 31, 2017, <https://www.daxx.com/article/software-developer-statistics-2017-programmers>