# HPC: History, State, Perspectives

## ICS632: Principles of High Performance Computing

Henri Casanova (henric@hawaii.edu)

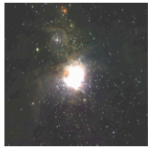
Fall 2015

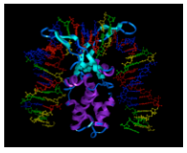
# Outline

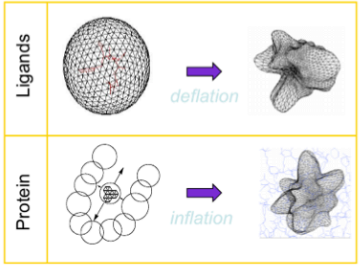
# "Killer" Apps: Big Computation, Big Data

## Science

- Global climate modeling
- Astrophysical modeling
- Biology (genomics; protein folding; drug design)
- Computational Chemistry
- Computational Material Sciences and Nanosciences

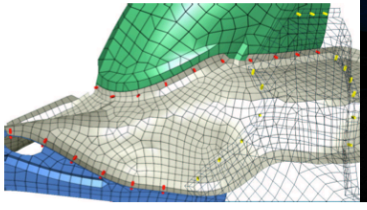


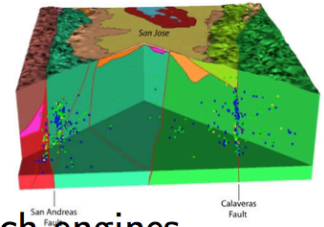


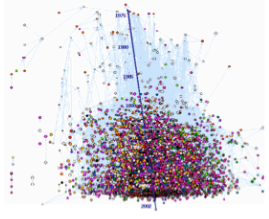



## Engineering

- Crash simulation
- Semiconductor design
- Earthquake and structural modeling
- Computation fluid dynamics (airplane design)
- Combustion (engine design)





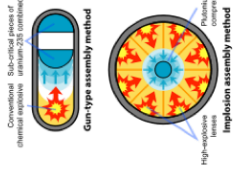



## Business and Humanities

- Financial and Economic modeling
- Transaction processing, web services and search engines
- Social Networking

## Defense

- Nuclear weapons – tested by simulations
- Cryptography

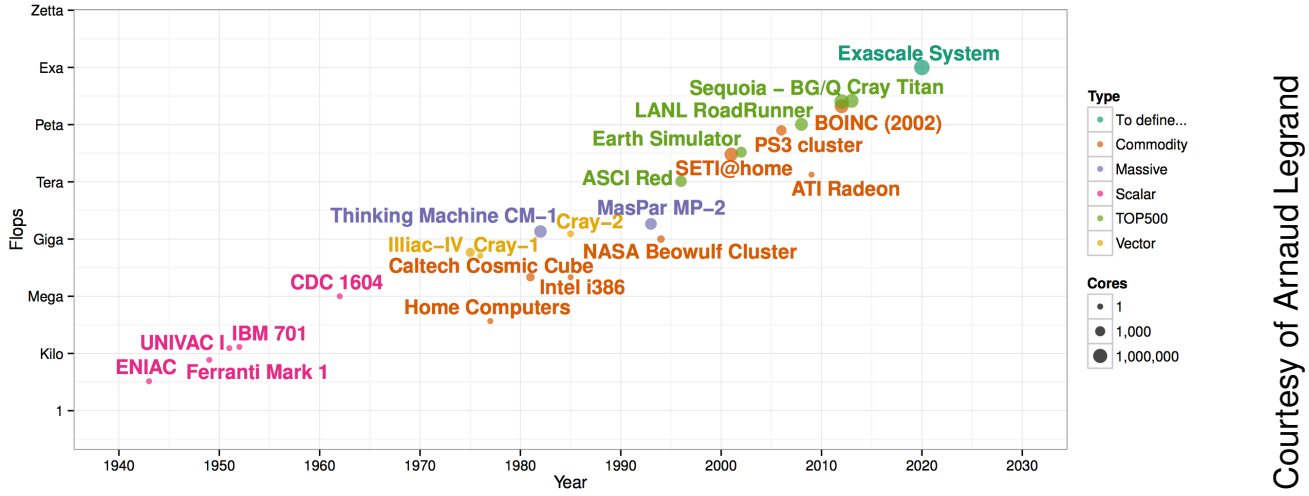



Courtesy of Martin Quinson (2011)

# Science and Supercomputing

- Scientific applications are large (arbitrarily?)
- Scientific applications are interesting/important
- We want results as quickly as possible to modify/invalidate scientific theories, engineering processes, etc.
- The biggest machines historically have been the ones designed particularly for running scientific application
- The so-called "Supercomputers", a vague term that implies:
  - A large number of state-of-the-art compute nodes
  - A fast network

# A Bit of History

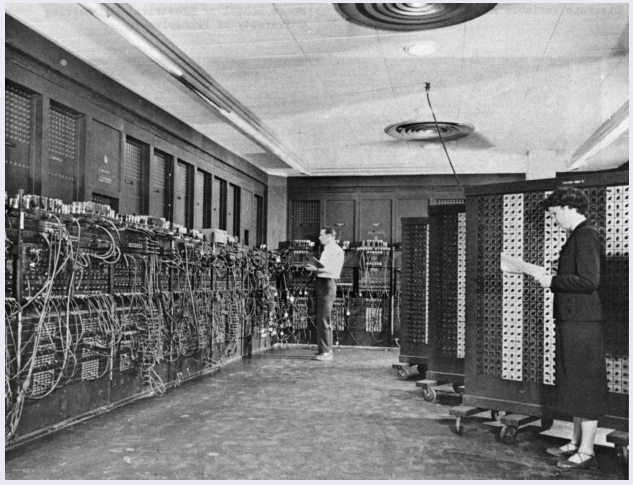



Courtesy of Arnaud Legrand

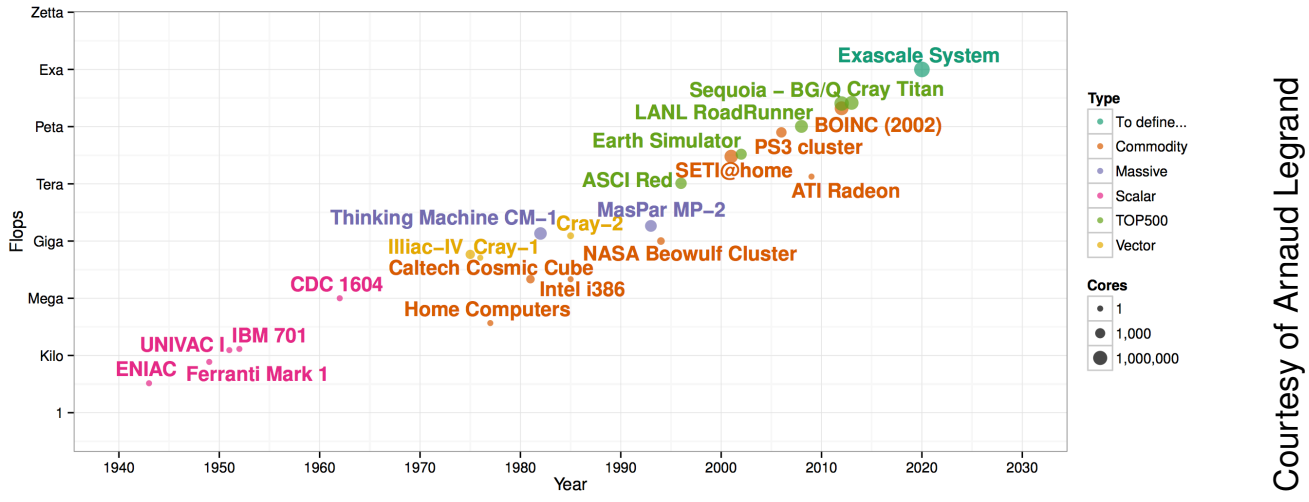
## 1943: The early days

ENIAC: 35 Flops ($\sim$ \$6,000,000)

"It was possible to connect several accumulators to run simultaneously, so the peak speed of operation was potentially much higher due to parallel operation."



# A Bit of History



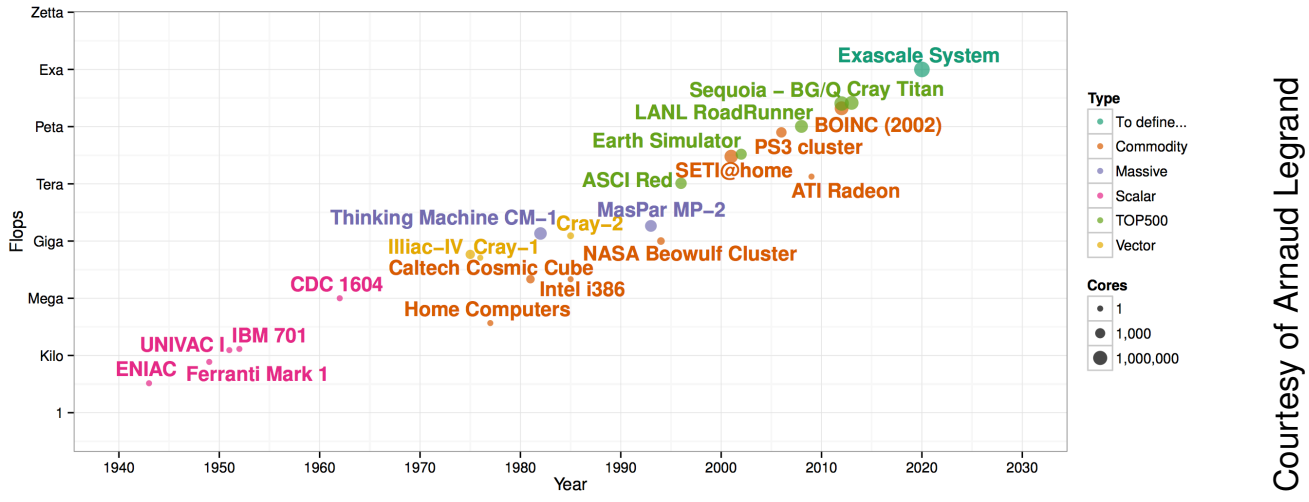

Courtesy of Arnaud Legrand

## 1949: The early days

Manchester Mark 1

Ran a Mersenne Primes search for 9 hours without error (Optimized later by Alan Turing!)

# A Bit of History




Courtesy of Arnaud Legrand

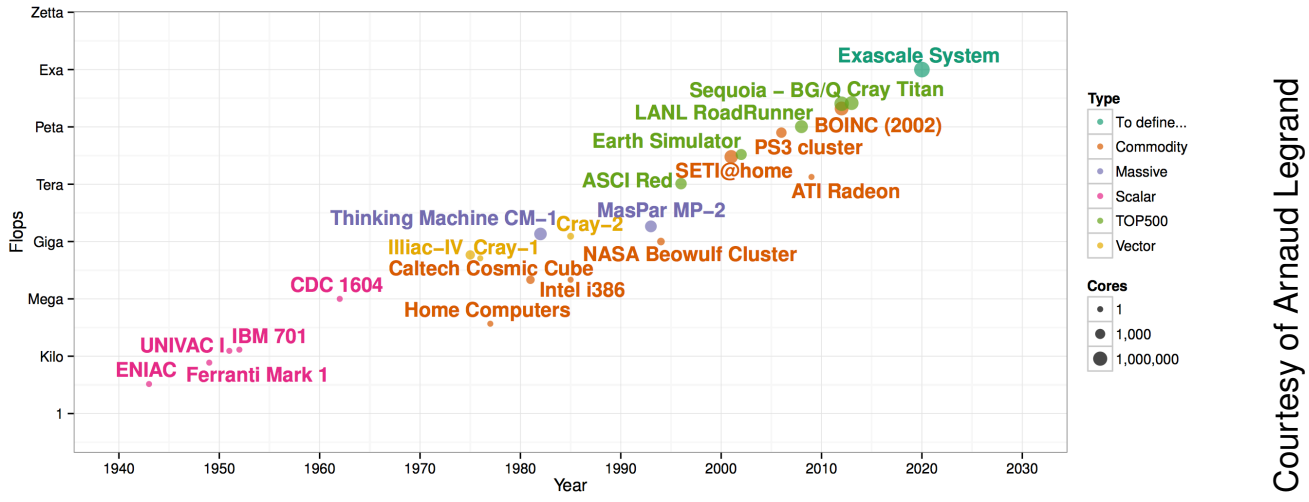
## 1951: A new market?

**Ferranti Mark 1**: First commercially available general-purpose electronic computer (460 Flops)

**UNIVAC 1 (Universal Automatic Computer)**: First "mass produced" computer

- The 5th machine (built for the U.S. Atomic Energy Commission) was used by CBS to predict the result of the 1952 presidential election
- 46 machines sold at more than $1 million each (equivalent to $8.95 million in 2012)

# A Bit of History




Courtesy of Arnaud Legrand

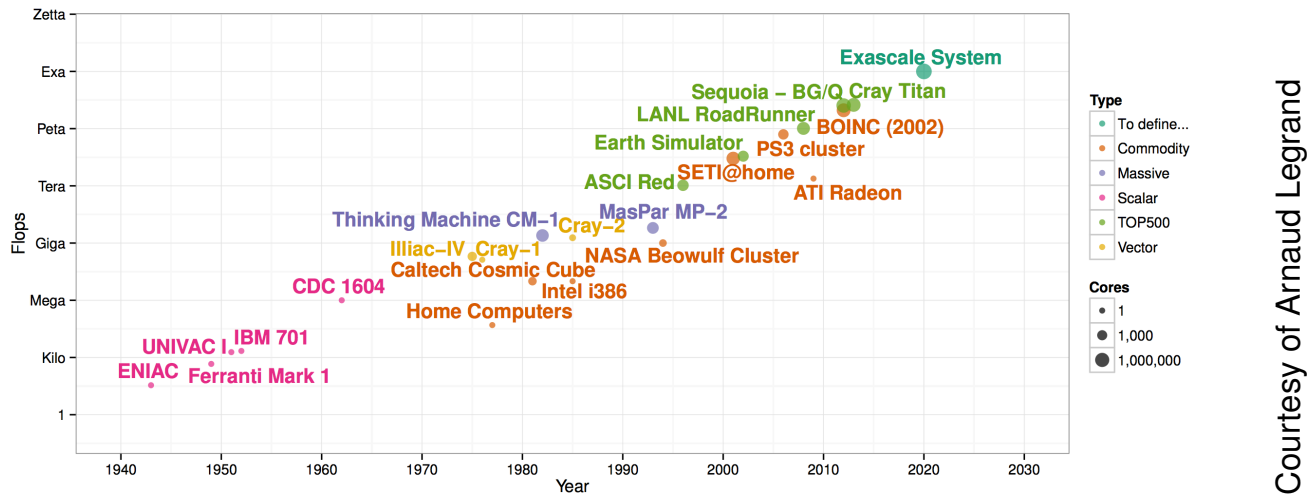
## 1952: A new market!

**IBM 701** (aka Defense Calculator) is IBM's's first commercial scientific computer. 2,200 FLOPS. Rental charge was about $12,000 a month.

"I think there is a world market for maybe five computers" – Thomas Watson Jr.

He visited 20 companies that were potential customers and said: *"as a result of our trip, on which we expected to get orders for five machines, we came home with orders for 18"*

# A Bit of History




Courtesy of Arnaud Legrand

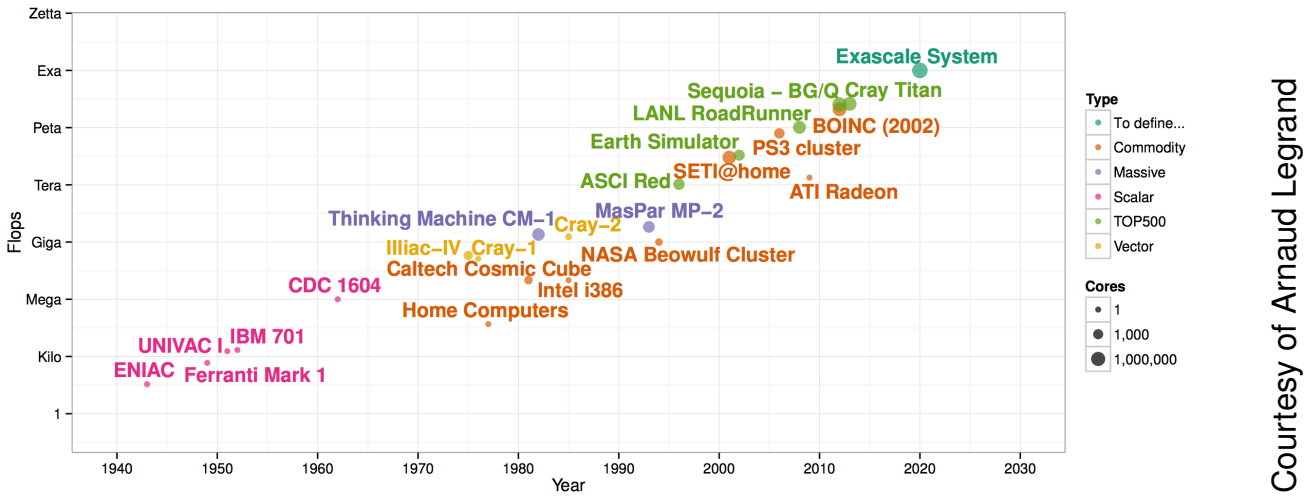
## 1962: Control Data Corporation

**CDC 1604**, first *transistor-based* computer, delivered to the US Navy (0.1 Mflops)

Designed by **Seymour Cray** and his team

One processor, 48-bit words, 6-$\mu$s memory

# A Bit of History




Courtesy of Arnaud Legrand

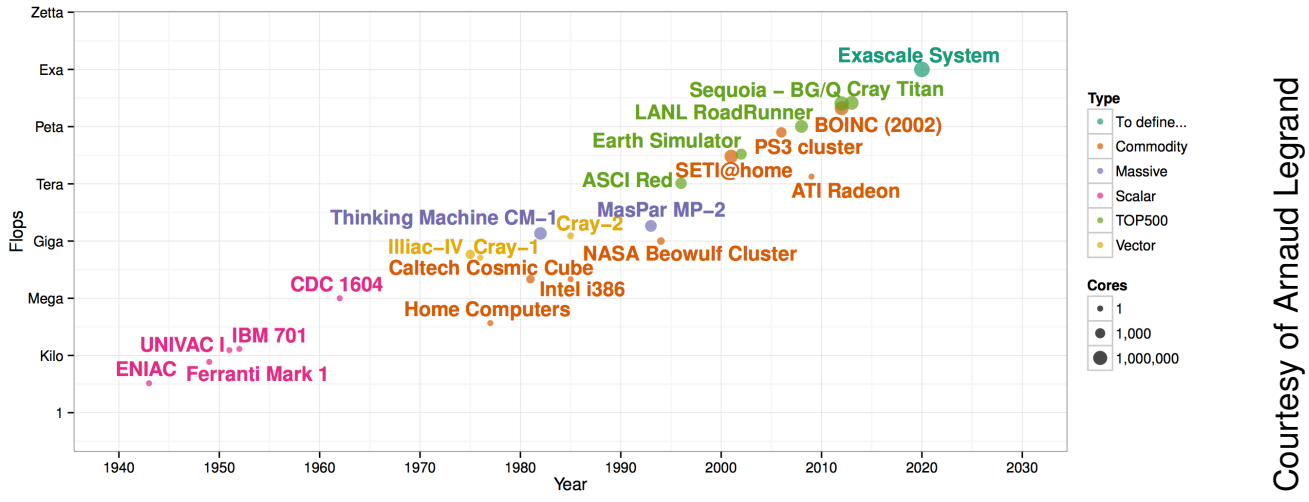
## 1966-1975: Illiac-IV

**Illiac-IV** for NASA: 256 4-bit processing elements

Expected 1 GFlops but "only" 200 MFlops

**Micro-computers** from 1970

# A Bit of History



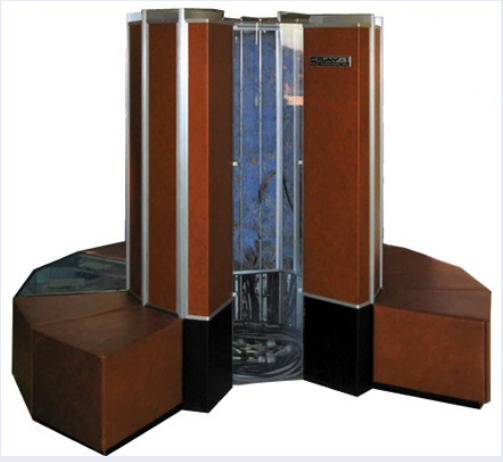

Courtesy of Arnaud Legrand

## 1976-1985: The Cray Era

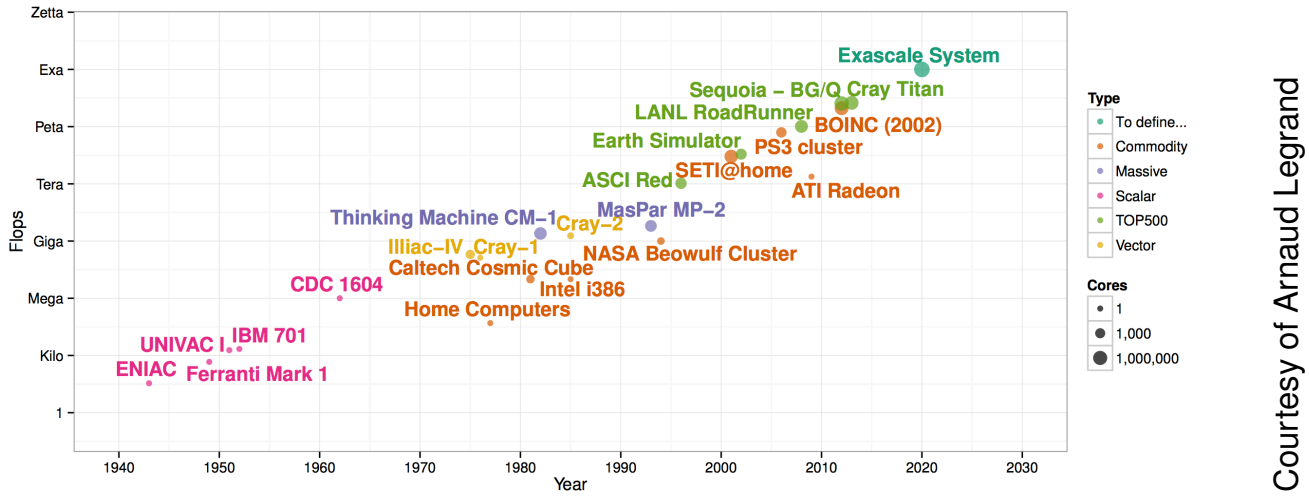
1976 **Cray-1:** Scalar + vector processor, 133 MFlops

1982 **Cray X-MP:** 800 MFlops with 2 to 4 CPUs

1985 **Cray-2:** 1,900 MFlops with 4 CPUs



# A Bit of History




Courtesy of Arnaud Legrand

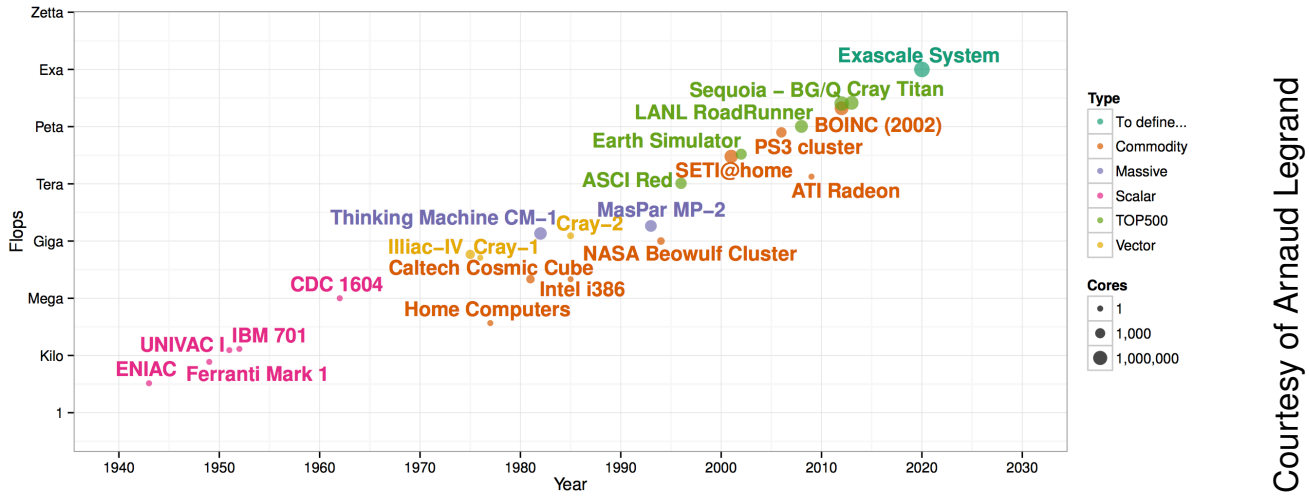
## 1976-1995: Massive Parallelism

1982 Thinking Machine's **CM-1**: 65,536 1-bit procs and a 12-D hypercube – 2,500 MFlops

1995 MasPar **MP-2**: 16,384 32-bit procs – 6,225 MFlops

1994-1997 **Cray T3D**: 128 procs – 19,200 MFlops

# A Bit of History



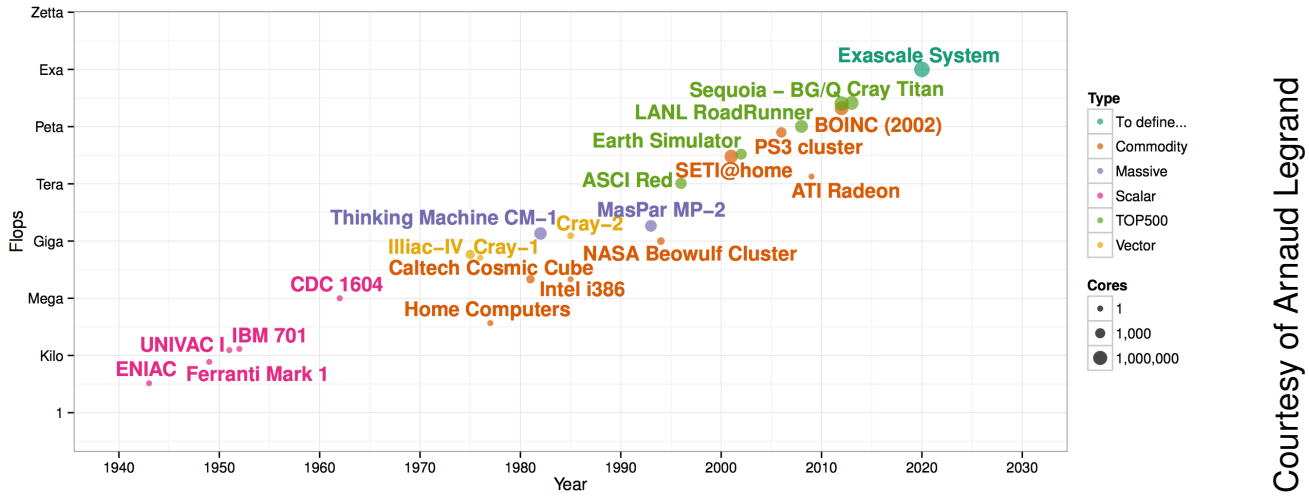

Courtesy of Arnaud Legrand

## 1981-1995: The Commodity Era

1981 **Caltech's Cosmic Cube:** 64-node hypercube with Intel 8086/8087 (10 MFlops)

1985 **Intel iPSC:** Commercial uptake

# A Bit of History

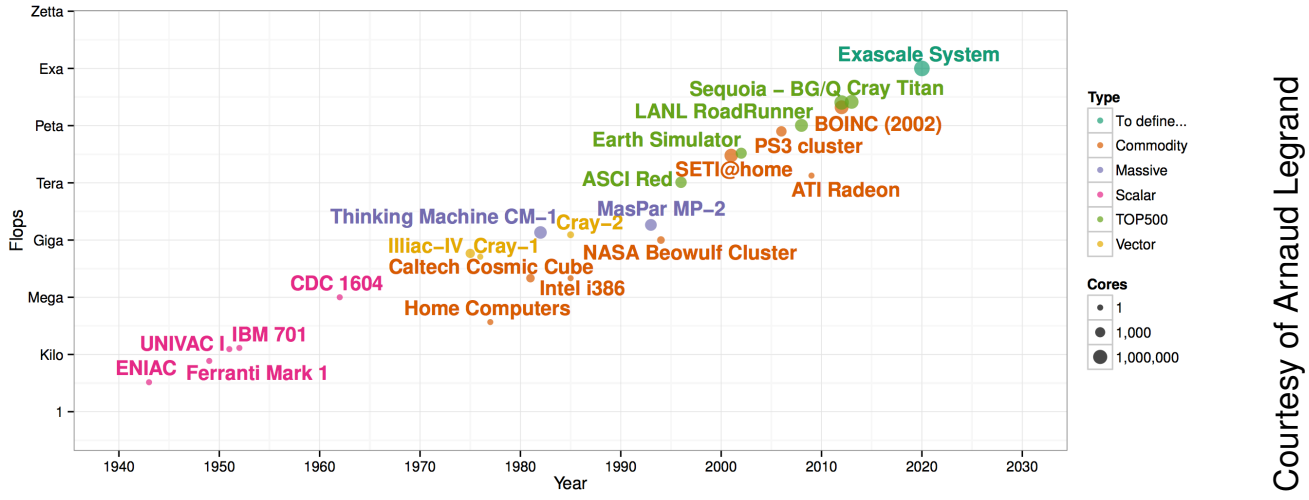



Courtesy of Arnaud Legrand

## 1981-1995: The Commodity Era

- Rack-mounted servers
- Blade servers

# A Bit of History



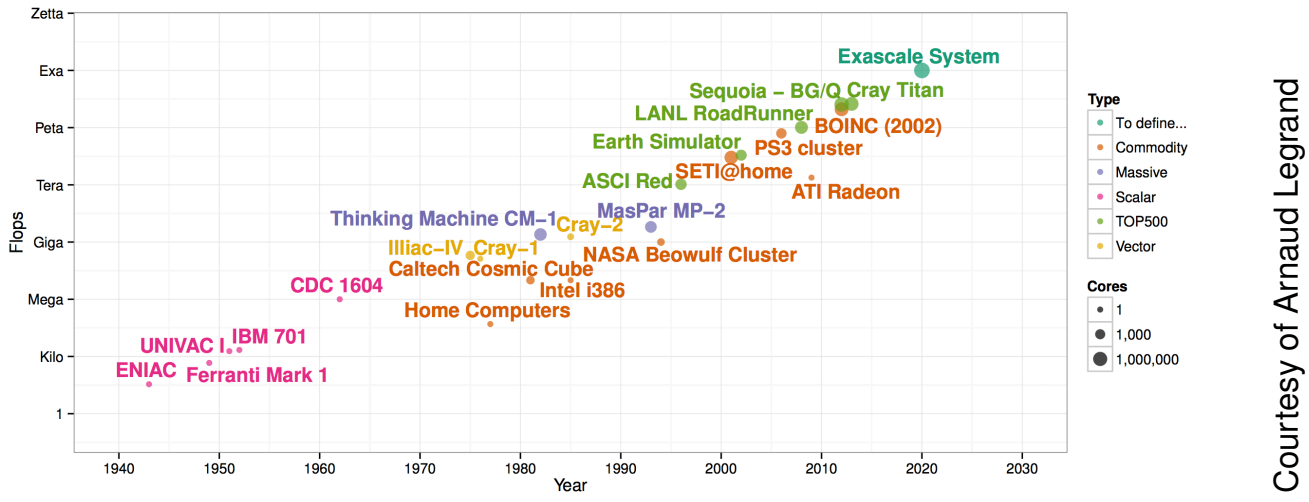

Courtesy of Arnaud Legrand

## 1981-...: The Commodity Era

1994 **Beowulf:** 16 PCs with Ethernet running Linux

- 1 GFlop for $50K
- "Do it yourself" attitude

# A Bit of History




Courtesy of Arnaud Legrand

## 1996-...: Massive Distributed Computing

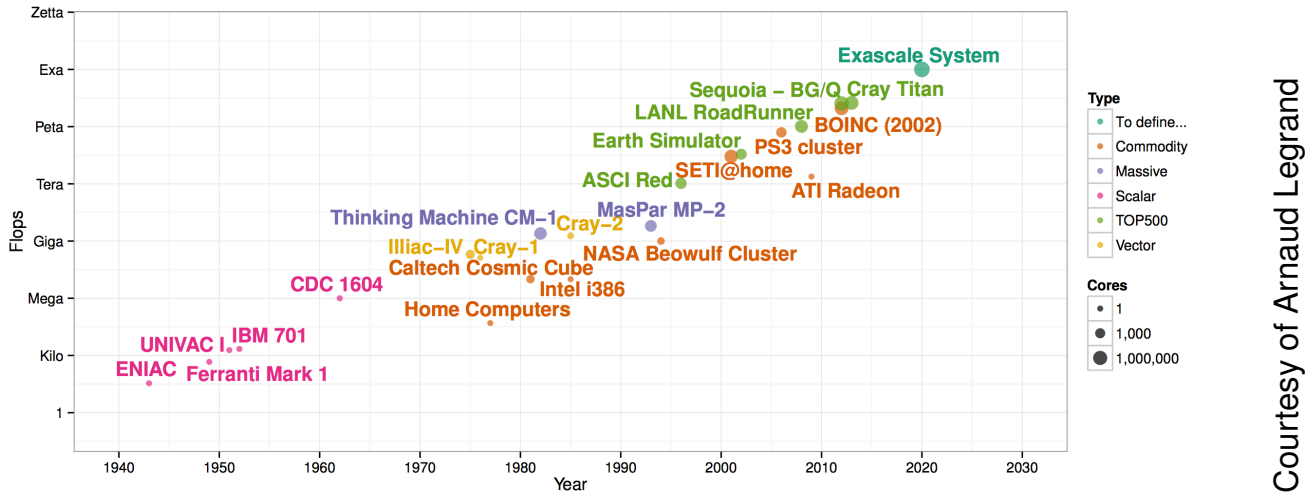
1999 **SETI@home**: 27.32 TFlops in 2002 with 300,000 hosts

since 2000 **Folding@home**, **LHC@home**, **DrugDiscovery@home**,...

2002 **BOINC**:

- Infrastructure for hosting projects
- 9.2PFlops in 2012 with 596,224 active hosts

# A Bit of History



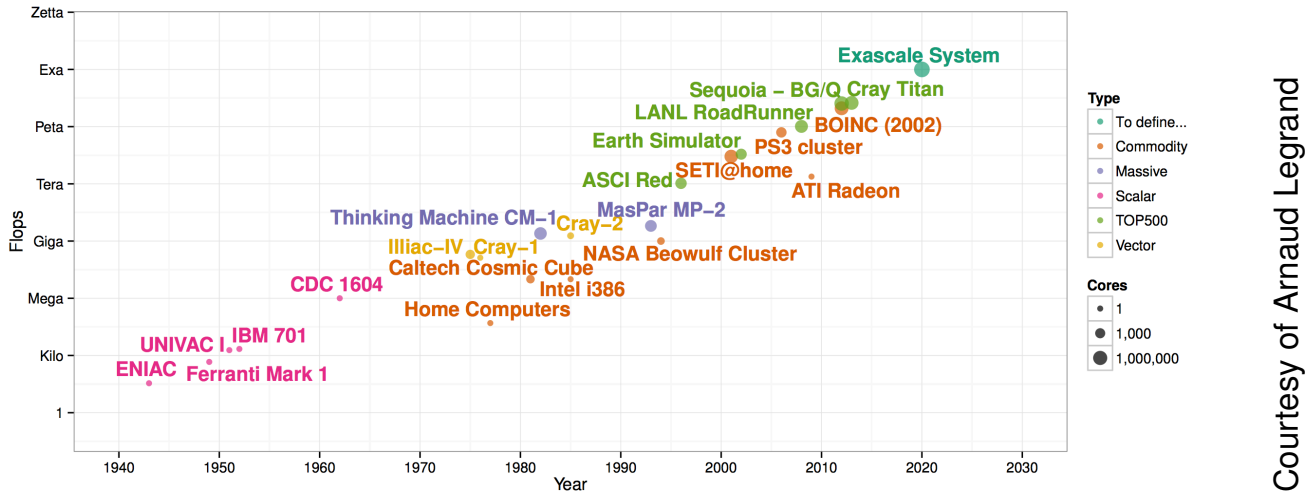

Courtesy of Arnaud Legrand

## 1996-...: "Commodity" Big Iron

1996-2001 **ASCI Red:** 1.06 TFlops, 9,298 Pentium Pro

2002 **Earth Simulator:** 35.9 TFlops, 640 nodes each with 8 vector processors

# A Bit of History




Courtesy of Arnaud Legrand

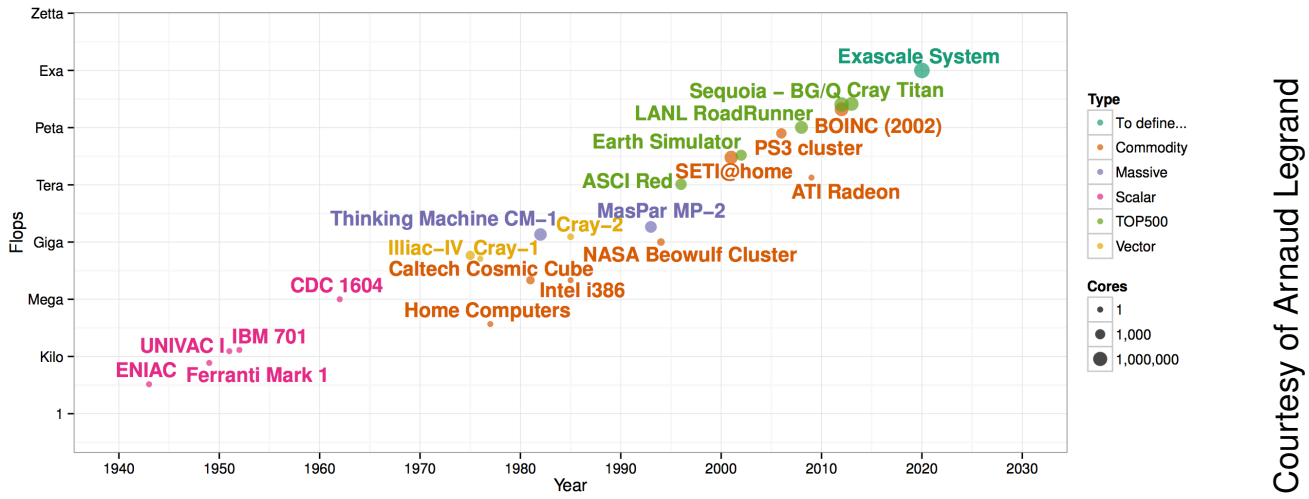
## 2001-...: Commodity Revolution

Computer entertainment market

GPGPU Computing

Today, $>$ 1 TFlop on a single card!

# A Bit of History

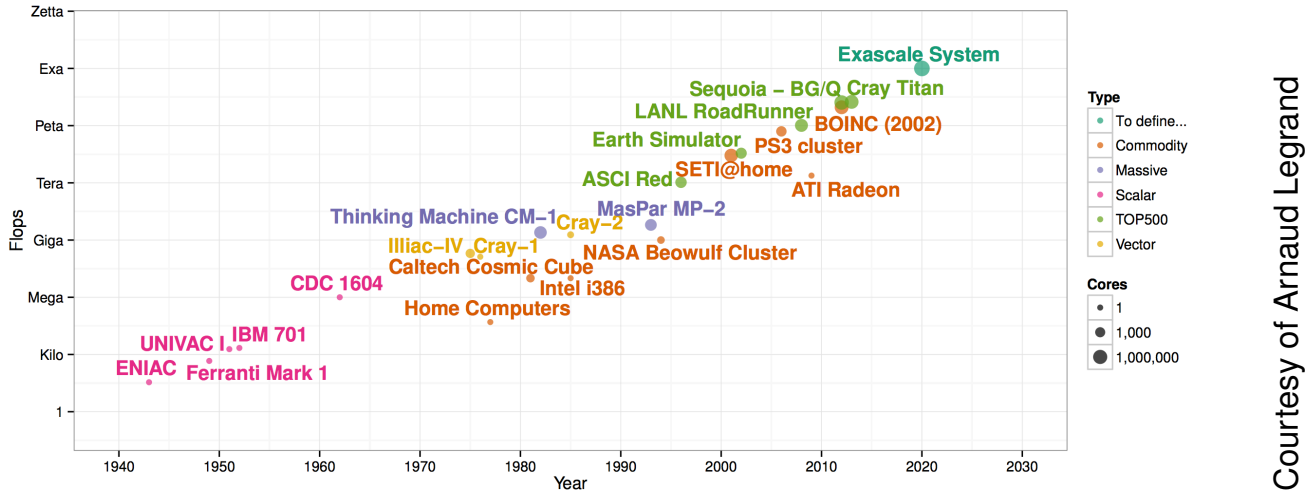



Courtesy of Arnaud Legrand

## 2000-...: Grid Computing

- An idea that's been overhyped, unhyped, renamed, reinvented, ... many times
- 2004-2011 **TeraGrid**

# A Bit of History

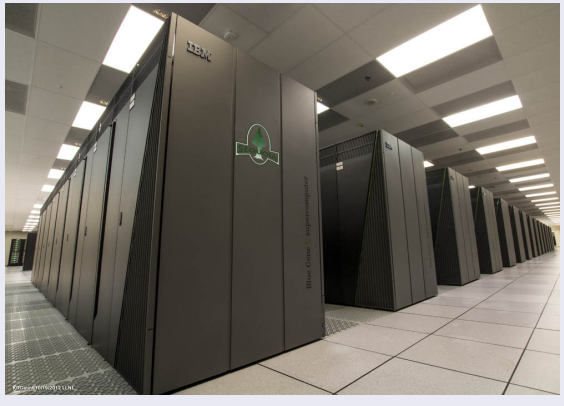



Courtesy of Arnaud Legrand

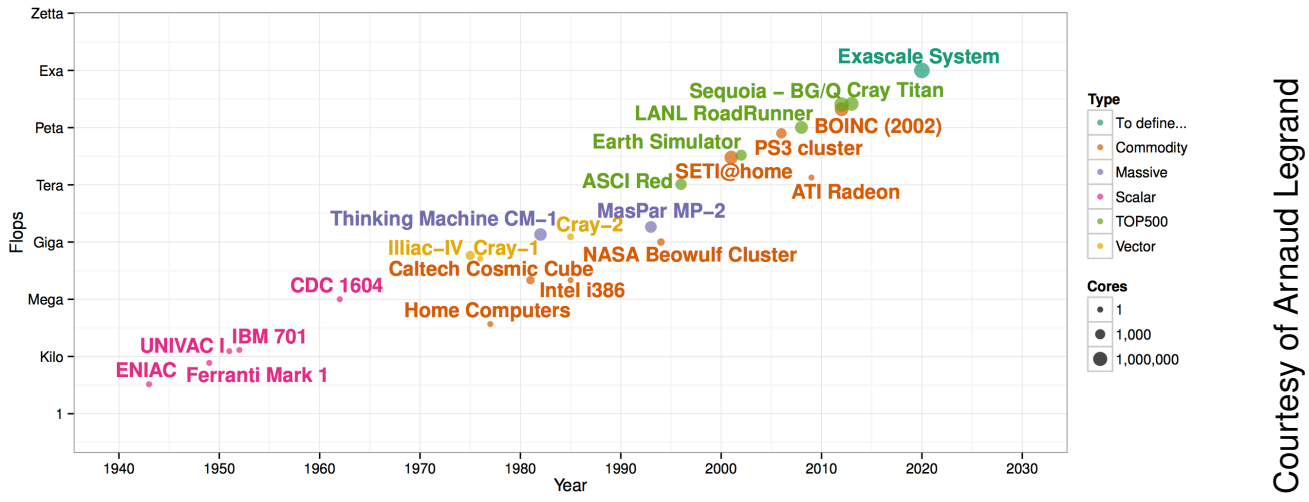
## 2012-2013: Petascale Systems

2012 **Sequoia (BlueGene/Q):** 98,304 16-core PowerPC processors (1,572,864 cores), 16.32 PFlops, 7.9 MW

At SNL, used for Nuclear weapons simulation mainly but also astronomy, human genome, climate change, etc.



# A Bit of History

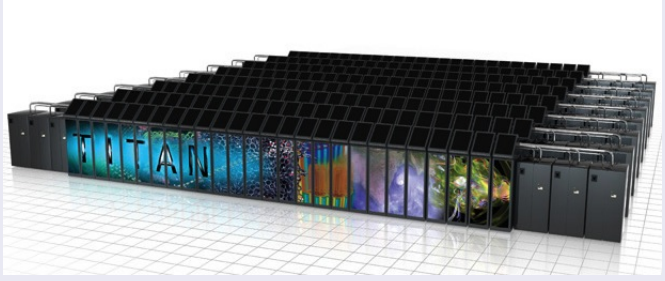



Courtesy of Arnaud Legrand

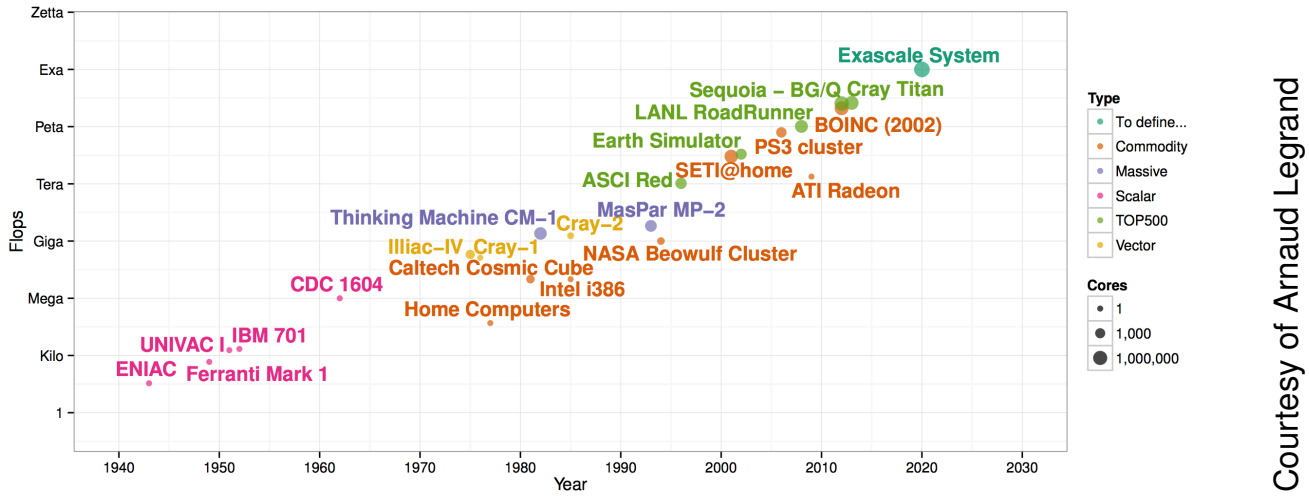
## 2012-2013: Petascale Systems

2013 **Titan (Cray):** 562,960 AMD cores + Nvidia GPUs, 17.59 PFlops, 8.2 MW

At ORNL, used for mostly for physics simulations



# A Bit of History




Courtesy of Arnaud Legrand

## 2012-2013: Petascale Systems

2013 **Tianhe-2:** 32,000 Ivy Bridge procs + 48,000 Xeon Phi procs (3,120,000 cores), 30.65 PFlops, 17.6 MW

At Sun Yatsen Univ., China, used for "simulation, analysis, and government security applications"

# An Incredible Evolution

## In 1996

- ASCI Red $\sim$ 1.24 GFlops / Watt
  - This Watt count does not include cooling

# An Incredible Evolution

## In 1996

- ASCI Red $\sim$ 1.24 GFlops / Watt
  - This Watt count does not include cooling

## Today

- Tesla K40 $\sim$ 18.25 GFlops / Watt

# Outline

# The Top500 List


www.top500.org

- List of 500 "fastest supercomputers" in the world
- Based on the Linpack benchmark (dense linear equations)
  - How representative of your workload?
- Ranking published every 6 months with technical specs
- Performance results:
  - $R_{max}$: Maximum Linpack performance
  - $R_{peak}$: Theoretical peak performance
  - Power: Electrical power consumption

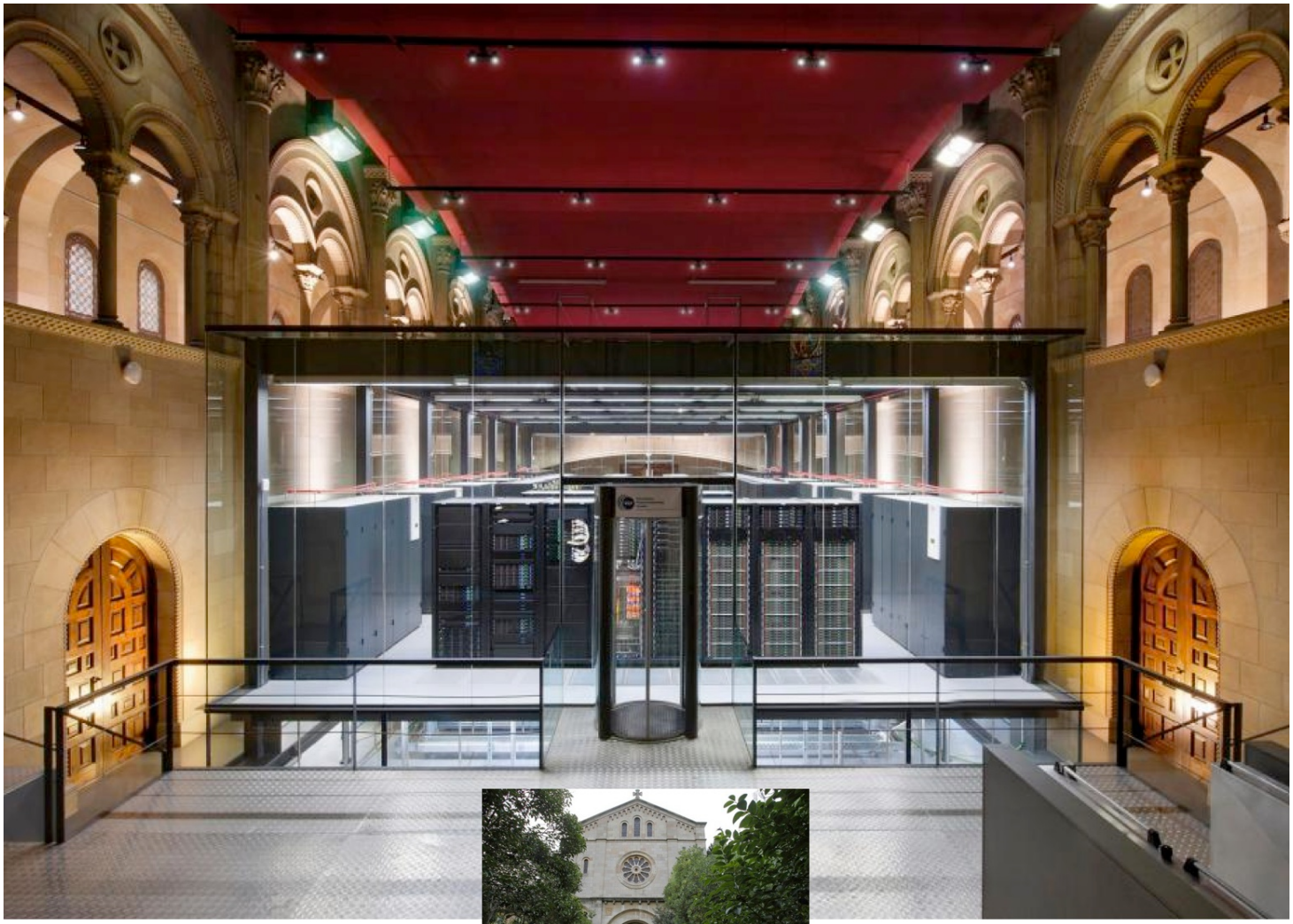
# Exploring Top500

- Clusters (and MPPs) dominate:
  - Statistics > List Statistics > Architectures
- Historical trends
  - Feature > Top500 Timeline
  - Statistics > Development over Time > Architectures
  - Statistics > Development over Time > Countries
  - Statistics > Development over Time > Operating Systems
  - Statistics > Development over Time > Cores per Socket

# Is our Cray cluster on the Top500?

- Not quite :(
- The 178 "standard" nodes are reported together to run the LINPACK benchmark at 72.6 TFlops
- The 6 "large memory" nodes are reported together to run the LINPACK benchmark at 4.6 TFlops
- Soon 50 standard nodes will be added
- Horrible guestimate: $72.6 \times (178 + 50)/178 + 4.6$, which gives 97.5 Tflops
- The last ranked system on the June 2015 Top500 is at 164.8 TFlops
  - Our system would have made the June 2013 list!!

# And our Cray isn't even in a church!



# Convergence

- Scientific computing and business computing have converged
- Data-centers are very much like HPC machines
  - Same interconnects, same hardware, increasingly similar workloads
- There is cross-pollination of ideas and technologies
  - VMs for science, accelerators for business
- Cloud computing and Grid computing are very related
- Amazon's cloud was #41 on Top500 in 2011!
  - In spite of VM overhead!!

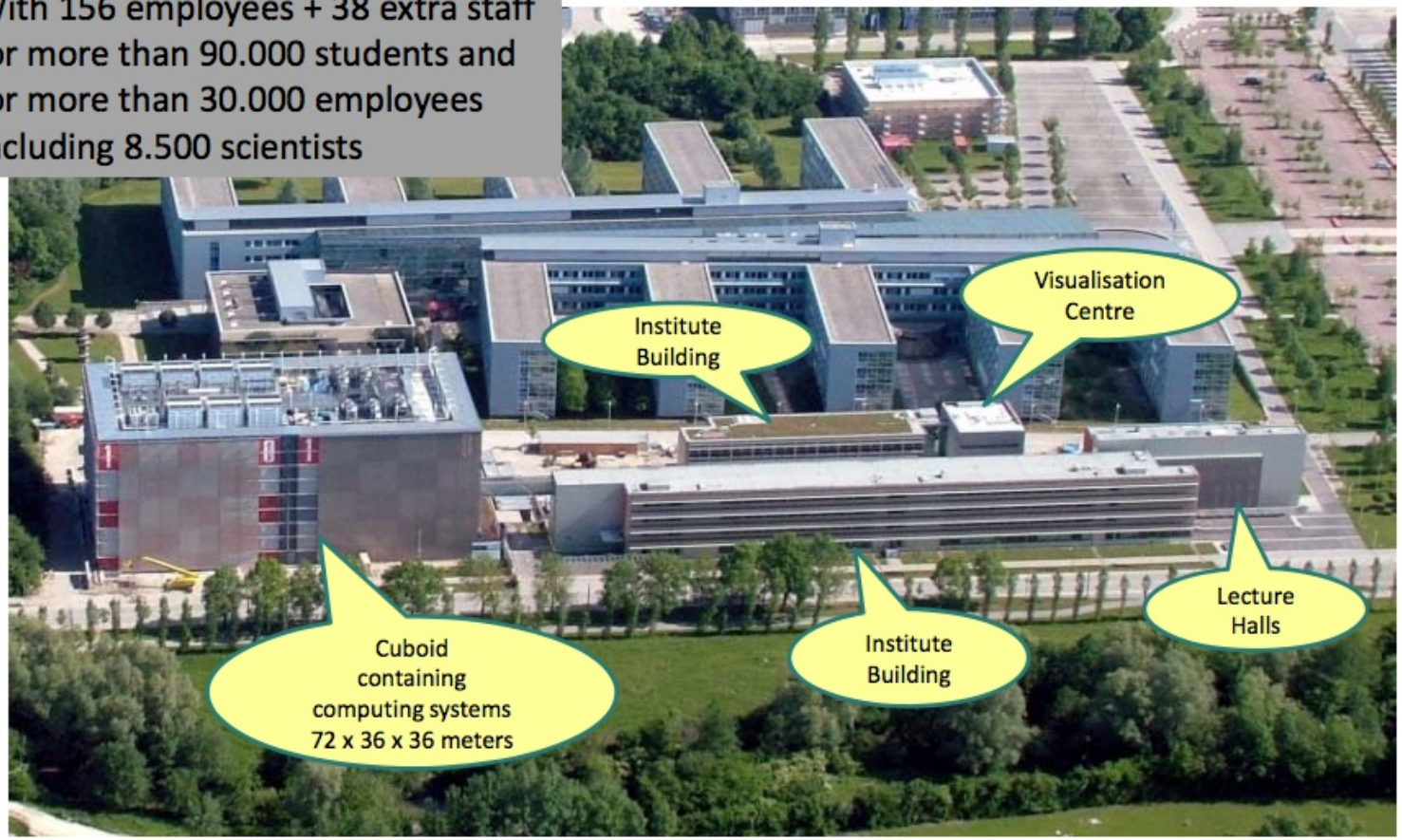
# Overview of a System

- Each time I teach this course, I show a few slides ~~stolen~~ borrowed from a recent presentation I happened to see at a conference/workshop
- At a Dagstuhl workshop, Prof. Dieter Kranzmüller (München, Germany) talked about SuperMUC
  - IBM system at the Leibniz Supercomputing Center (München)
  - Two-part system for over 200K cores
  - Currently ranked #20 and #21 on Top500 (was up to #4)
- Let's look at some of these slides (just pretty pictures)...

# Leibniz Supercomputing Center

# Leibniz Supercomputing Center

With 156 employees + 38 extra staff for more than 90.000 students and for more than 30.000 employees including 8.500 scientists



# The Cuboids (one being built a few years ago)


Figure: Herzog+Partner für StBAM2 (staatl. Hochbauamt München 2)

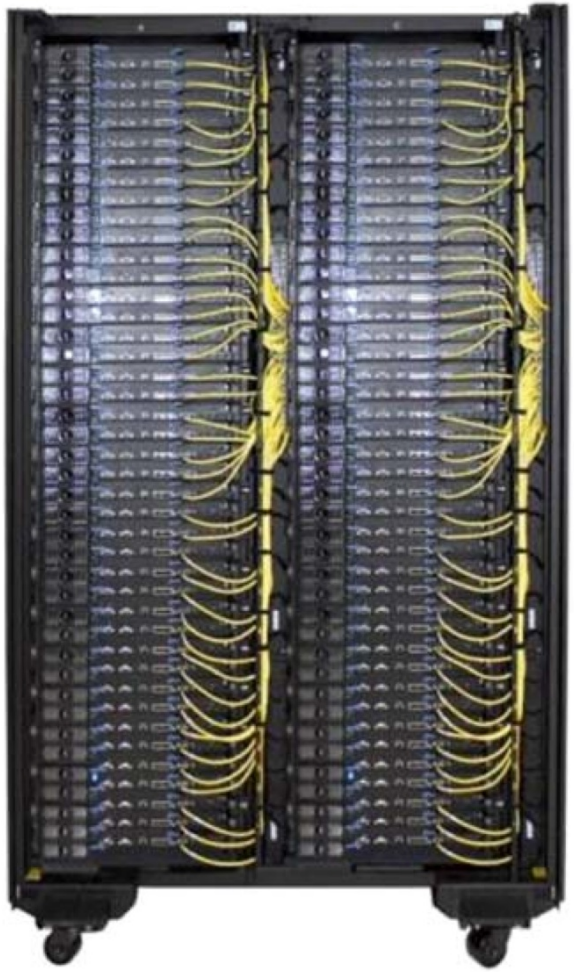
# The Cuboids from Google Streetview

# Water Cisterms and Pumps

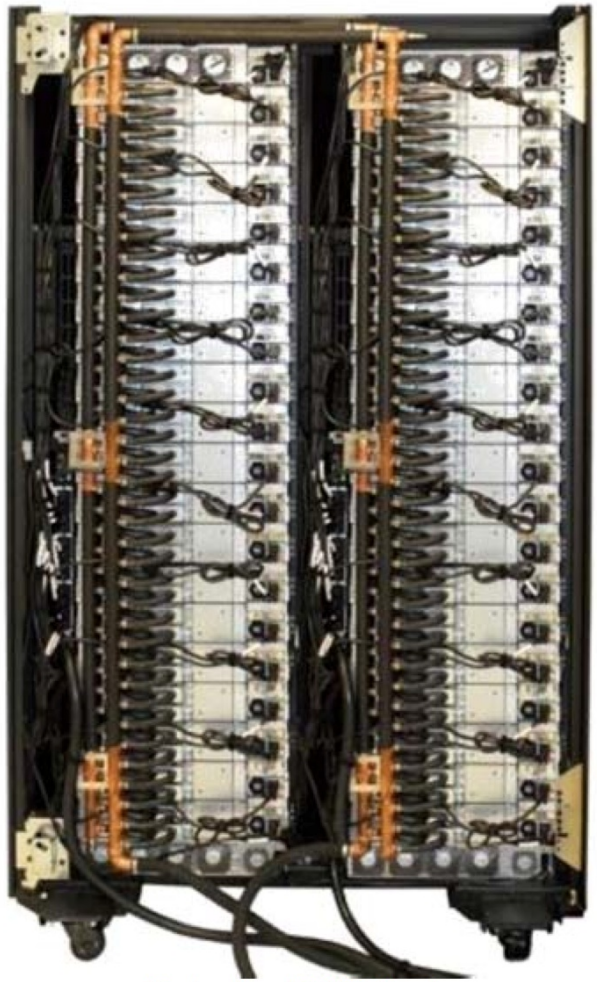
**Water is pumped from the river and stored in cisterns that can deliver cool water quickly**

# Water up to the Machine Room Floor

# Direct Water Cooling to the Cabinets

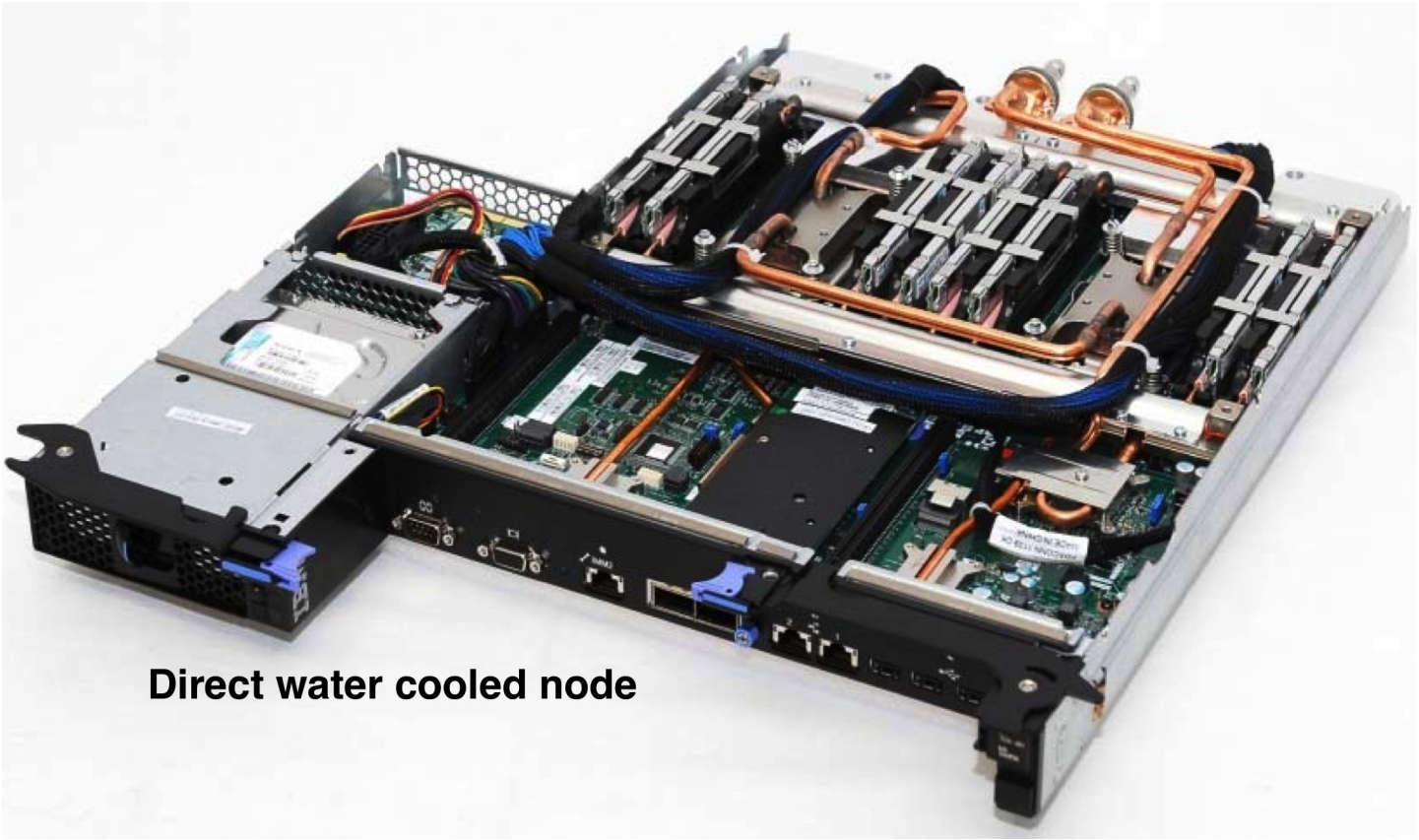


iDataplex DWC Rack
w/ water cooled nodes
(front view)



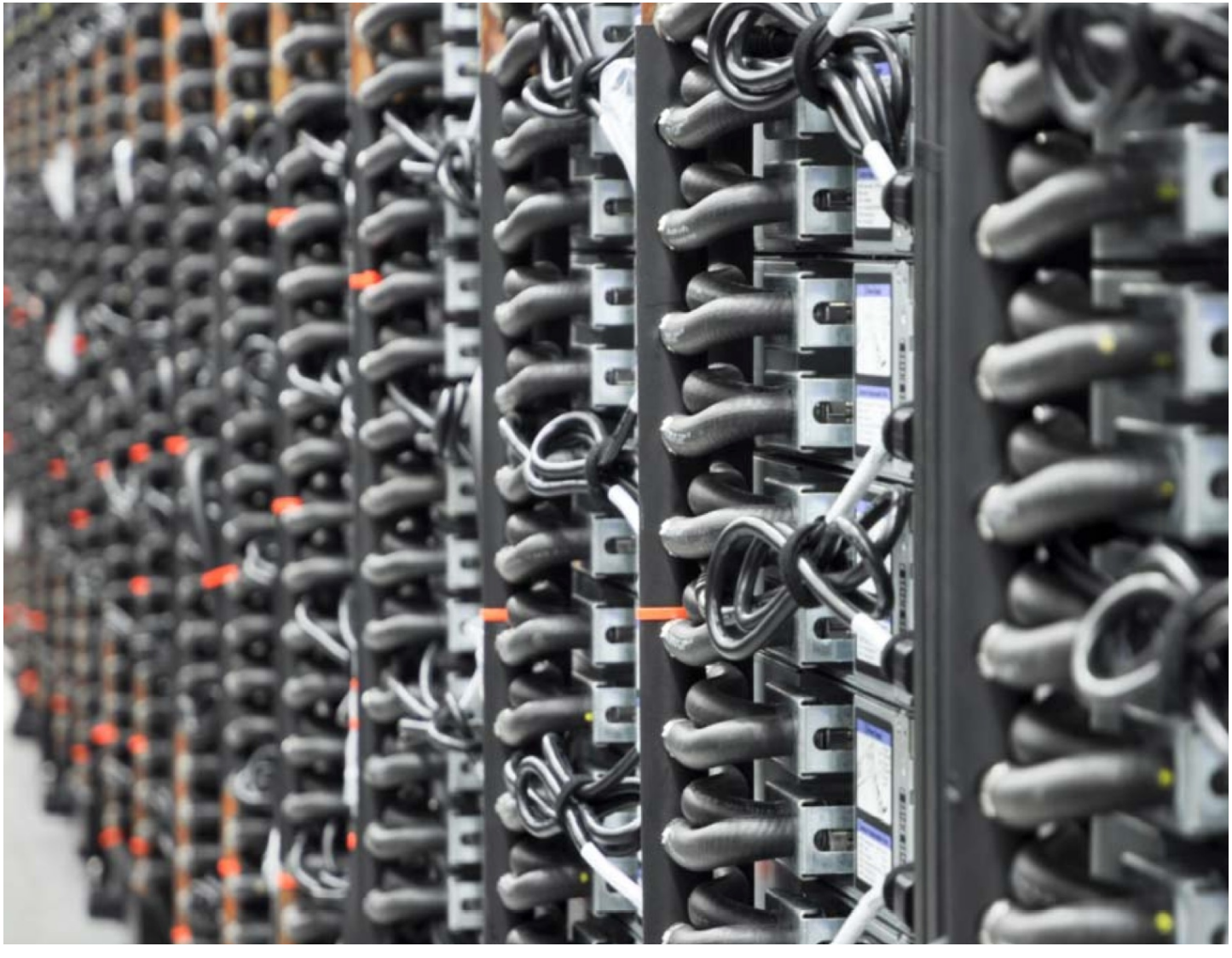
iDataplex DWC Rack
w/ water cooled nodes
(rear view of water manifolds)

# Direct Water Cooled Compute Nodes

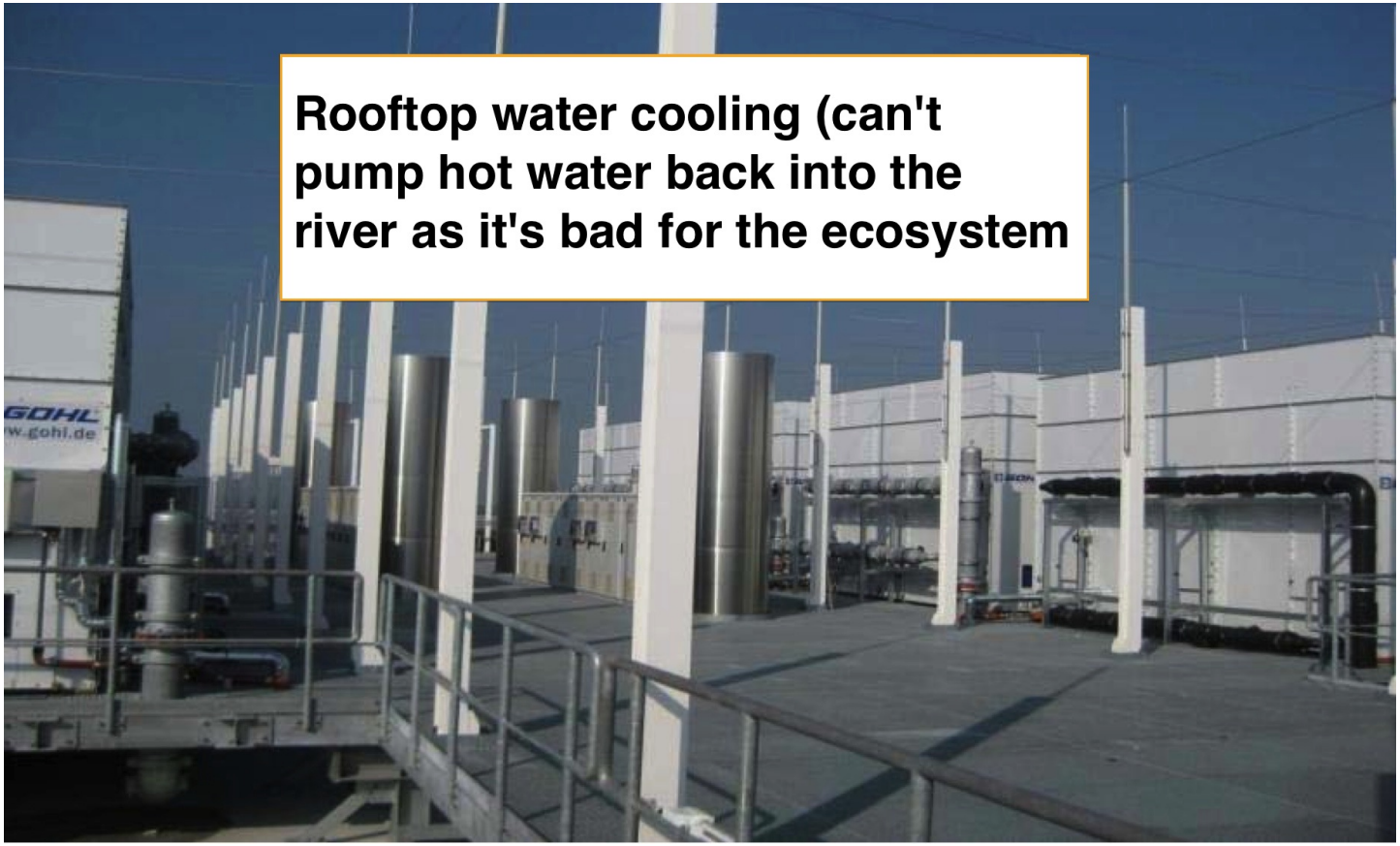


Direct water cooled node

# Tons of Water Hoses and Gaskets (leaks???)



# Rooftop Water Chilling



# SuperMUC

- This is a pretty cool (not super new) machine
- The direct water cooling is a great way to save on energy, compared to A/C cooling, which is used in most systems
  - Our cray uses watercooled radiators mounted on the cabinets to cool off the hot air produced by the nodes so as to reduce A/C usage
- SuperMUC was designed to be very energy efficient
  - Now it's "only" ranked #85 on the Green500 list
- We could spend the semester looking at HPC systems
  - Some of my research collaborators are living encyclopedias of HPC coolness in terms of hardware and infrastructure
- We'll consider these issues a bit in the course, but will focus more on "how to use the machine?" rather than on "how to build the machine?"

# Outline

# Exascale Systems

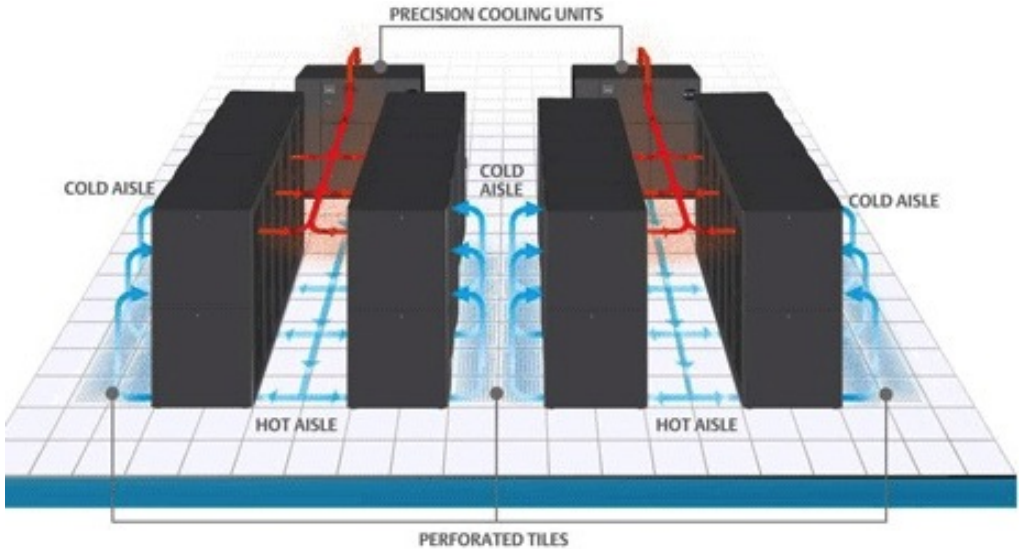
- 1 Exaflops ($10^{18}$ flops) expected before 2020
  - Top500 $>$ Statistics $>$ Performance Development
- Raises many challenges
  - Power consumption / Cooling
  - Performance scalability and programmability
  - Reliability
  - ...
- Let's look a DoE's challenge list here
- A very popular research area with many conferences, workshops, etc.
  - Each new "<prefix>scale" is the next exciting frontier
  - Some question whether we should attempt Exascale

# Exascale and Power

- Main challenge: Electrical Power
- Assuming a 20MW power budget (Tianhe-2 is at 17.8MW), we would need a 50 GFlops/Watt
  - Tianhe-2 is at about 2 GFlops/Watt
  - http://www.green500.org
- GPUs can make this power efficiency much higher
  - Take care of peak vs. real performance!
  - Double Precision arithmetic is slower
  - Programming model is different
  - ...
- ARM processors can be used as well!
  - Mont-Blanc project (ARM + GPU)

# Exascale Heat Challenge

- There is ongoing research for reducing the power consumption of all components of the system
- Processors consume almost half of the power budget, network, storage and memory consume a large fraction
  - Solutions: new technology, turn off when idle, DVFS, etc.
- There rest is consumed by cooling
  - Processors below threshold temperature (e.g., below 85 C)




Courtesy of http://www.datacenterknowledge.com

# Exascale Fault-Tolerance Challenge

- Compute nodes have a finite Mean Time Between Failure (MTBF), on order of decades, say $125$ years
- But with, say, with $2^{20}$ compute nodes, there would be a failure on average every hour
  - Tianhe-2 has $2^{18}$ compute nodes
- This is a problem for running large applications
- Current systems already suffer from frequent failures
- Basic idea: save state periodically, or *checkpointing*
  - Problem: too much overhead, kills parallel efficiency
  - Solutions: replications, failure predictions, better checkpointing, etc.

# Conclusion

- The history of "supercomputing" is long and interesting
- We have gone from a world with machines custom-built on demand to a world of commodity hardware and software
  - There are vendors of these commodity hardware/software systems, and they just buy components from Intel, AMD, etc.
- GPUs have caused a lot of upheaval and are (or other accelerators like the Xeon Phi, etc.) are likely to increase their presence in the Top500 lists