Fra superdatamaskiner til grafikkprosessorer og maskinlæring

Brødtekst

Prof. Anne C. Elster

IDI HPC/Lab

Parallel Computing: Personal perspective

- 1980's: Concurrent and Parallel Pascal
- 1986: Intel iPSC Hypercube
  - CMI (Bergen) and Cornell
  (Cray arrived at NTNU)
- 1987: Cluster of 4 IBM 3090s
- 1988-91: Intel hypercubes
  - Some on BBN
- 1991-94: KSR (MPI1 & 2)

Intel iPSC

Kendall Square Research (KSR)
KSR-1 at Cornell University:

- 128 processors – **Total RAM: 1GB!!**
- Scalable shared memory multiprocessors (SSMMs)
- Proprietary 64-bit processors

Notable Attributes:

Network latency across the bridge prevented viable scalability beyond 128 processors.

The World is Parallel!!

All major processor are now multicore chips!

--> All computer devices and systems are parallel

… even your Smartphone!

WHY IS THIS?

Why is computing so exciting today?

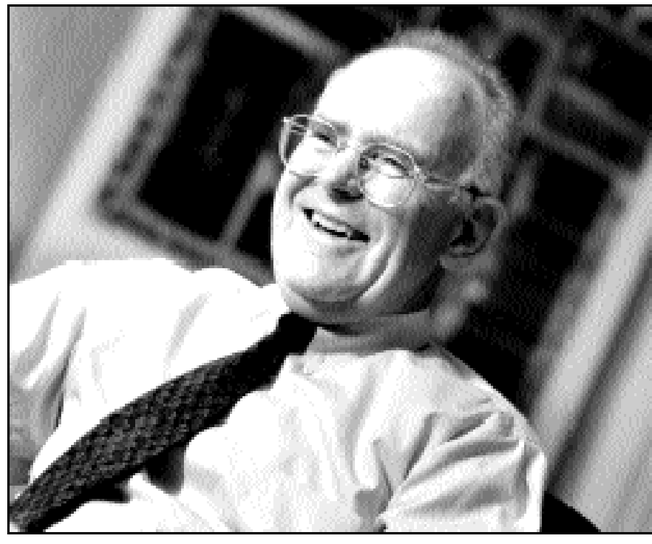
- Look at the tech. trends!

Microprocessors have become smaller, denser, and more powerful.

As of 2016, the commercially available processor with the highest number of transistors is the 24-core Xeon Haswell-EX with > 5.7 billion transistors. (source: WikiPedia)

NVIDIA

Tech. Trend: Moore's Law

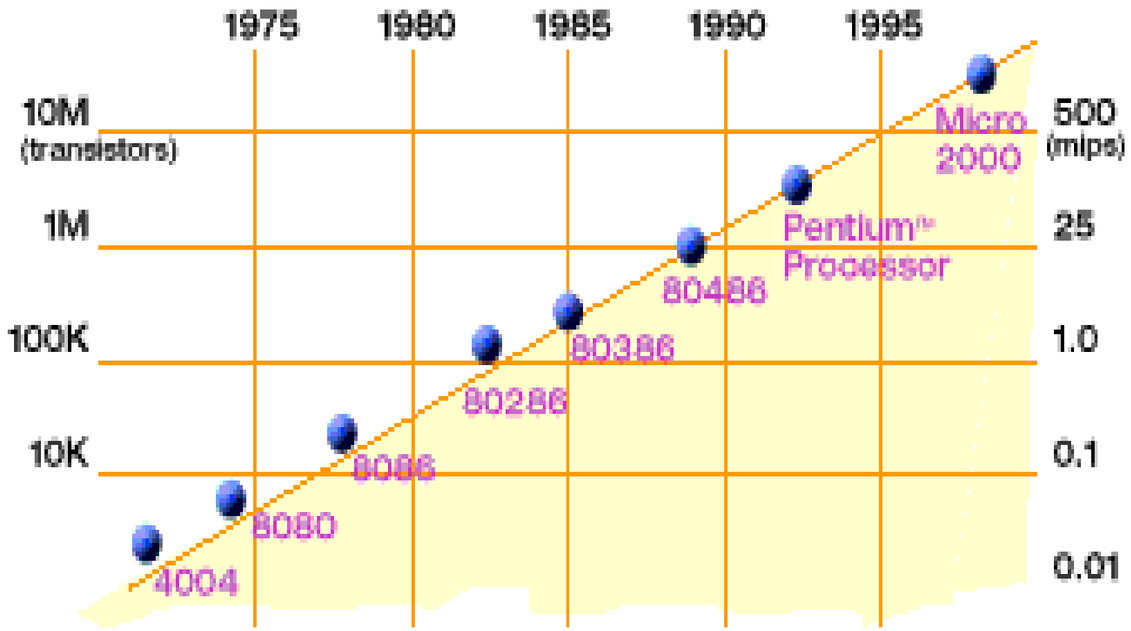


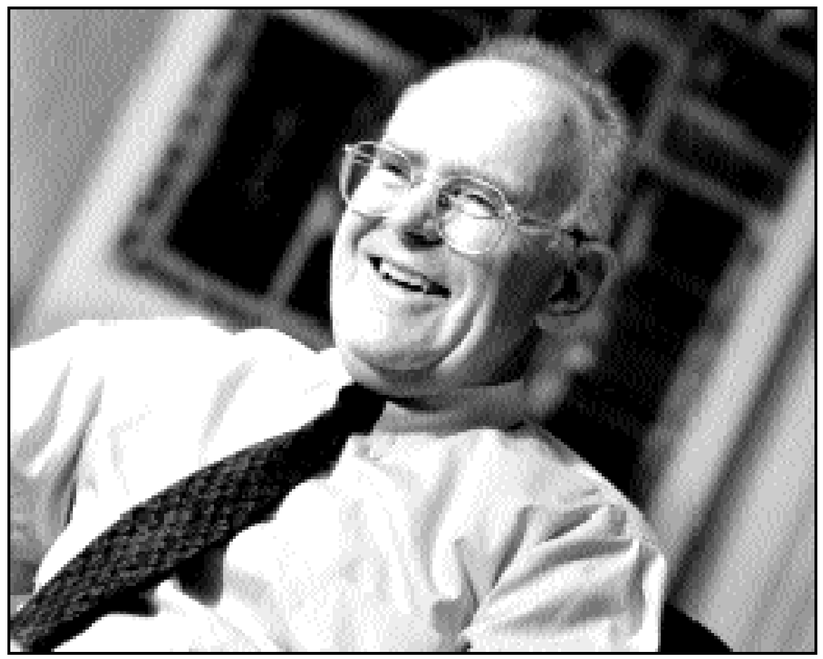
- **Named after Gordon Moore (co-founder of Intel)**
- **Moore predicted in 1965 transistor density of semiconductor chips would double roughly every year, revised in 1975 to every 2 years by 1980**
- **Some think is says that it actually doubles every 18 months since use more transistors and each transistor is faster [due to quote by David House (Intel Exec)]**

"Moore's law" (popularized by Carver Mead, CalTech) is known as the observation and prediction that the number of transistors on a chip has and will be doubled approximately every 2 years.

But in 2015: Intel stated that this has slowed starting in 2012 (22nm), so now every 2.5 yrs (14nm (2014), 10nm scheduled in late 2017)

Tech. Trends: Microprocessor





2X transistors/Chip Every 1.5 years
Called "Moore's Law"

Microprocessors have become smaller, denser, and more powerful.

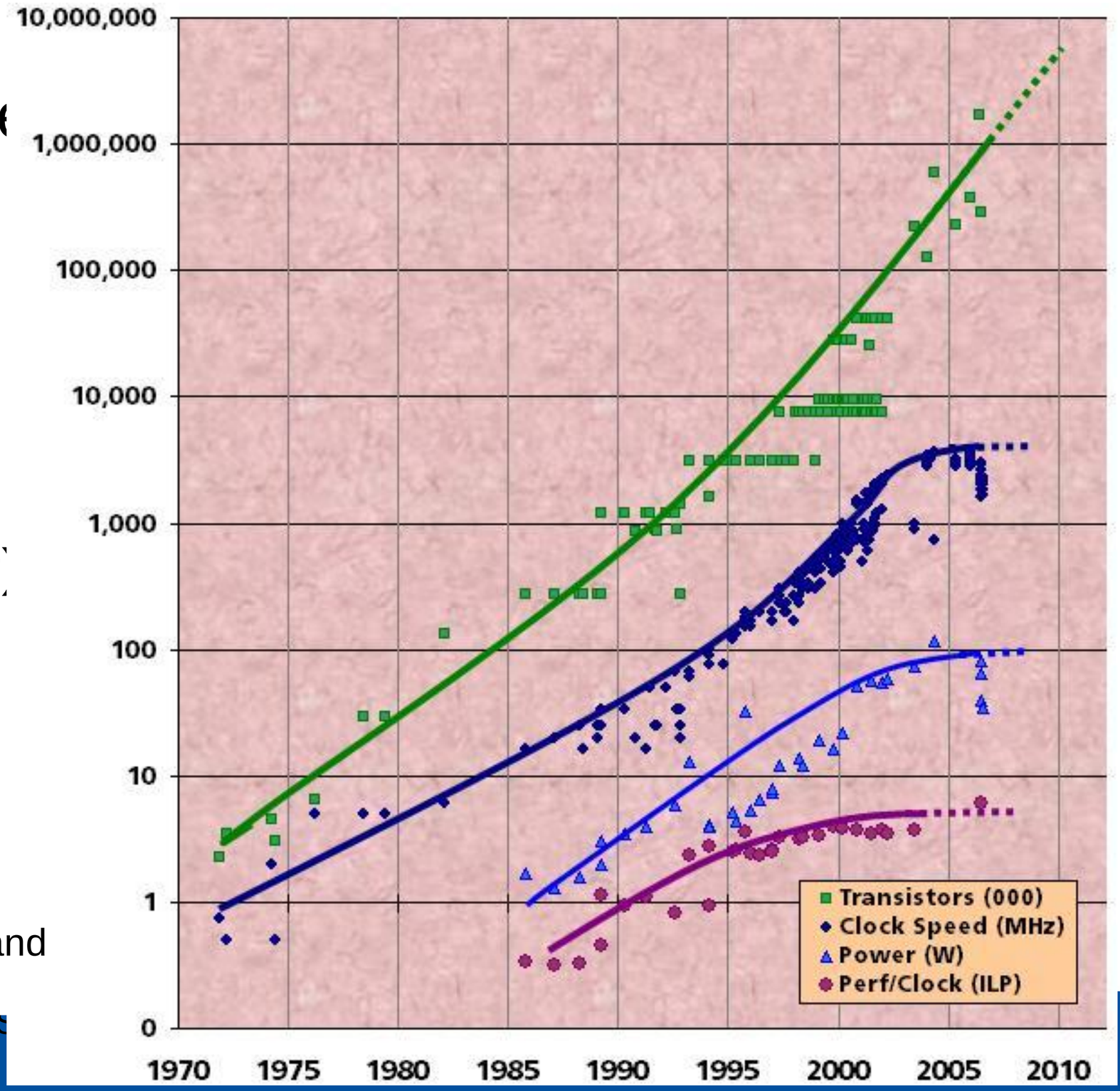
Gordon Moore (co-founder of Intel) predicted in 1965 that the transistor density of semiconductor chips would double roughly every 18 months.

01/17/2007 from CS267-Lecture 1

Slide source: Jack Dongarra

Revolution is Happening Now

- Chip density is continuing increase ~2x every 2 years
  - Clock speed is not
  - Number of processor cores may double instead
- There is little or no hidden parallelism (ILP) to be found
- Parallelism must be exposed to and managed by software

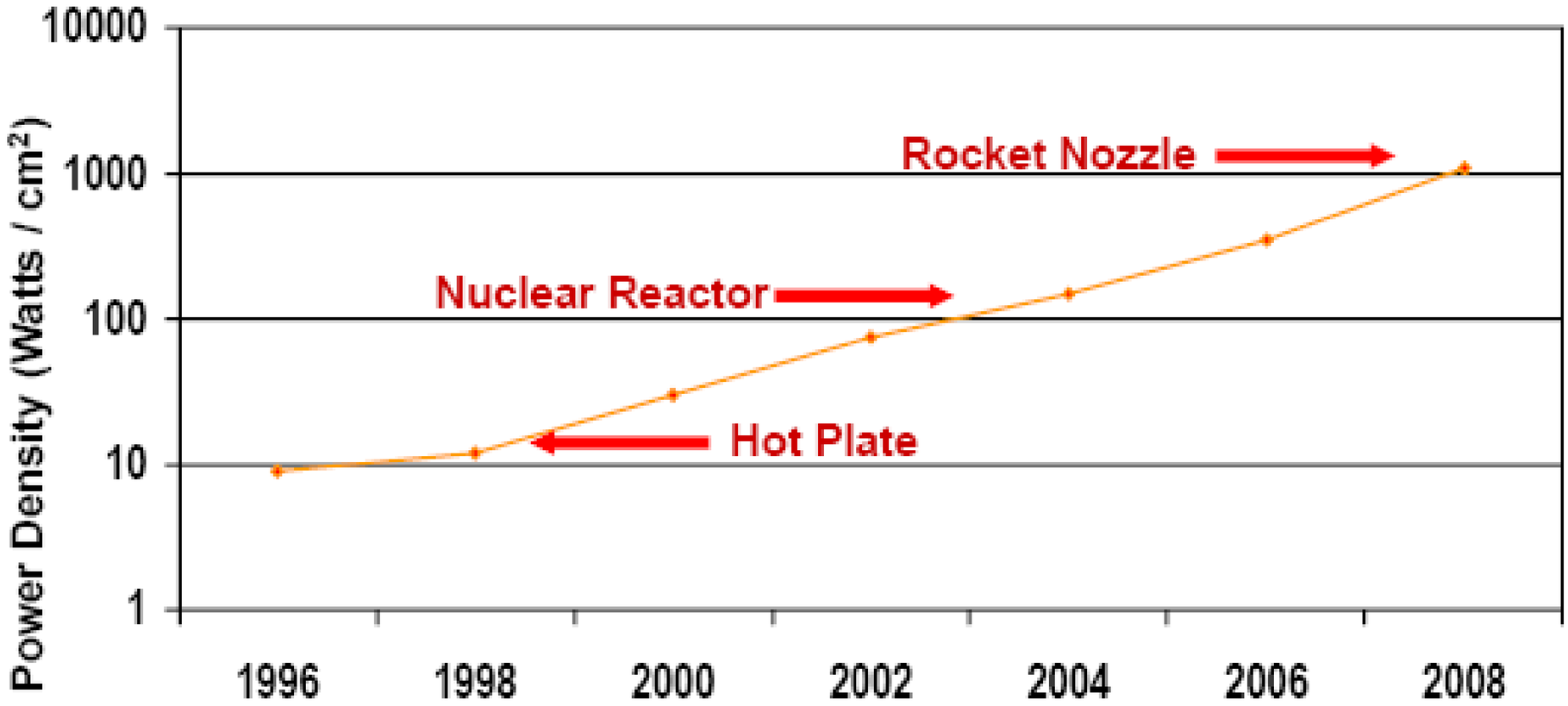


Source: Intel, Microsoft (Sutter) and Stanford (Olukotun, Hammond)

Power Density Limits Serial Performance

Moore's Law Extrapolation:

Power Density for Leading Edge Microprocessors



Power Density Becomes Too High to Cool Chips Inexpensively

Source: Shekhar Borkar, Intel Corp

What to do?

To increase processor performance one can:

1. Increase the system clock speed -> Power Wall(*)
2. Increase memory bandwidth-> more complex
3. Parallelize -> more complex

(*) The Power Wall:

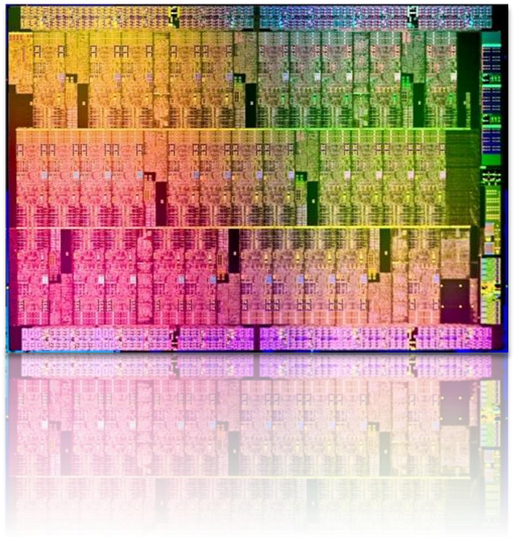
Too much heat and transistor performance degrades (more power leakage as power increases)!

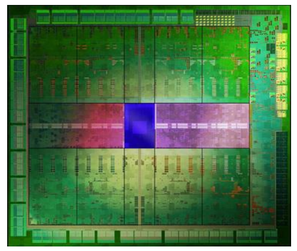
→ Now maxing out clock at 3-4GHz for general processors

Supercomputer & HPC Trends:
Clusters and Accelerators!

How did we get here?

NTNU IDI's 40-node AMD 1.46GHz Cluster w/ 2GB RAM, 40GB disk, Fast Ethernet – ca 2002-2003





Market forces!!

→ Rapid architecture development driven by gaming (graphics cards) and embedded systems architectures (e.g. ARM)

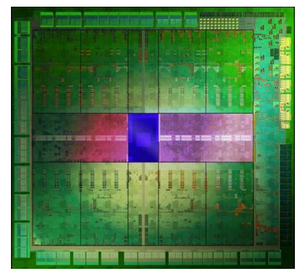
387 CUDA Teaching & Research Centers as of Aug 27, 2015!

Motivation – GPU Computing:

Many advances in processor designs
are driven by Billion $$ gaming market!
Modern GPUs (Graphic Processing Unit) offer
lots of FLOPS per watt!

NVIDA GTX 1080
(Pascal): 3640 CUDA cores!

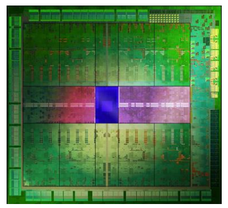
.. and lots of parallelism!



-Kepler:
-GTX 690 and Tesla K10 cards
-have 3072 (2x1536) cores!

TK1/Kepler

- GPU: SMX Kepler: 192 core
- CPU: ARM Cortex A15
  - 32-bit, 2instr/cycle, in-order
  - 15GBs, LPDDR3, 28nm process
- GTX 690 and Tesla K10 cards have 3072 (2x1536) cores!
- Tesla K80 is 2,5x faster than K10
  - 5.6 TF TFLOPs single prec.
  - 1.87 TFLOPS Double prec.
- Nested kernel calls
- Hyper Q allowing up to 32 simultaneous MPI tasks



TX1/Maxwell

- GPU: SMX Maxwell: 256 cores
  - 1 TFLOPs/s
- CPU: ARM Cortex-A57
  - 64-bit, 3 instr/cycle, out-of-order
  - 25.6 GBs, LPDDR4, 20nm process
- Maxwell Titan with 3072 cores
- API and Libraries:
  - Open GL 4.4
  - CUDA 7.0
  - cuDNN 4.0

NTNU IDI HPC-Lab (last 10 yrs)

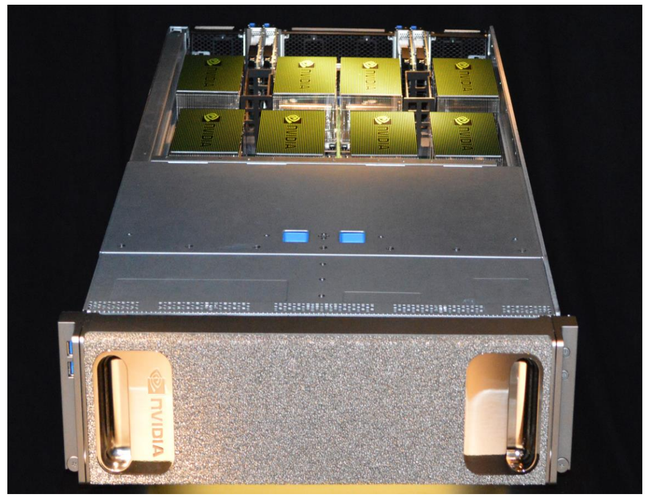
Fall 2006:

- First 2 student projects with GPU programming (Cg)

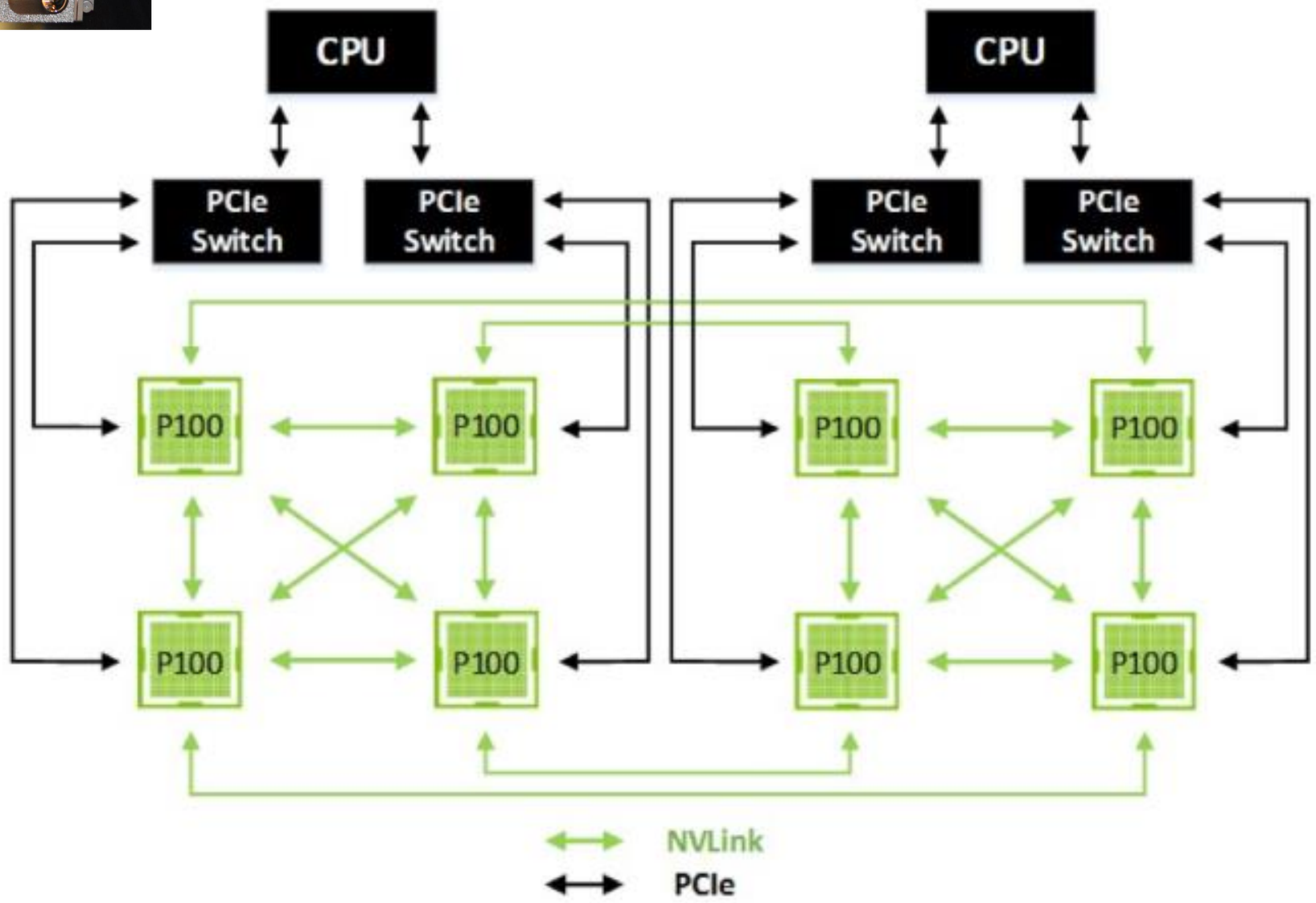
  Christian Larsen (MS Fall Project, December 2006):
  "Utilizing GPUs on Cluster Computers" (joint with Schlumberger)

Erik Axel Nielsen asks for FX 4800 card for project with GE Healthcare

Elster as head of Computational Science & Visualization program helped NTNU acquire new IBM Supercomputer
(Njord, 7+ TFLOPS, proprietary switch)



The NVIDIA DGX-1 Server



NVIDIA DGX-1 Server -- Details

CPUs : 2 x Intel Xeon E5-2698 v3 (16-core Haswell)

GPUs: 8 x NVIDIA Tesla P100 (3584 CUDA cores)

System Memory: 512 GB DDR4-23133

GPU Memory 128GB (8 x 16GB)

Storage: 4 x Samsung PM 863 1.9 TB SSD

Network: 4 x Infiniband EDR, 2x 10 GigE

Power¨: 3200W

Size 3U Blade

GPU Throughput: FP16: 170TFLOPs,

FP32: 85TFLOPs, FP 64: 42.5 TFLOPs

- Supercomputing / HPC units are:
  - Flop: floating point operation
  - Flops/s: floating point operations per second
  - Bytes: size of data (a double precision floating point number is 8)

- Typical sizes are millions, billions, trillions…

Mega	Mflop/s = 10^6 flop/sec	Mbyte = 2^{20} = 1048576 ~ 10^6 bytes
Giga	Gflop/s = 10^9 flop/sec	Gbyte = 2^{30} ~ 10^9 bytes
TeraTflop/s = 10^{12} flop/sec		Tbyte = 2^{40} ~ 10^{12} bytes
PetaPflop/s = 10^{15} flop/sec		Pbyte = 2^{50} ~ 10^{15} bytes
Exa	Eflop/s = 10^{18} flop/sec	Ebyte = 2^{60} ~ 10^{18} bytes
Zetta	Zflop/s = 10^{21} flop/sec	Zbyte = 2^{70} ~ 10^{21} bytes
Yotta	Yflop/s = 10^{24} flop/sec	Ybyte = 2^{80} ~ 10^{24} bytes

- See www.top500.org for current list of the world's fastest supercomputers