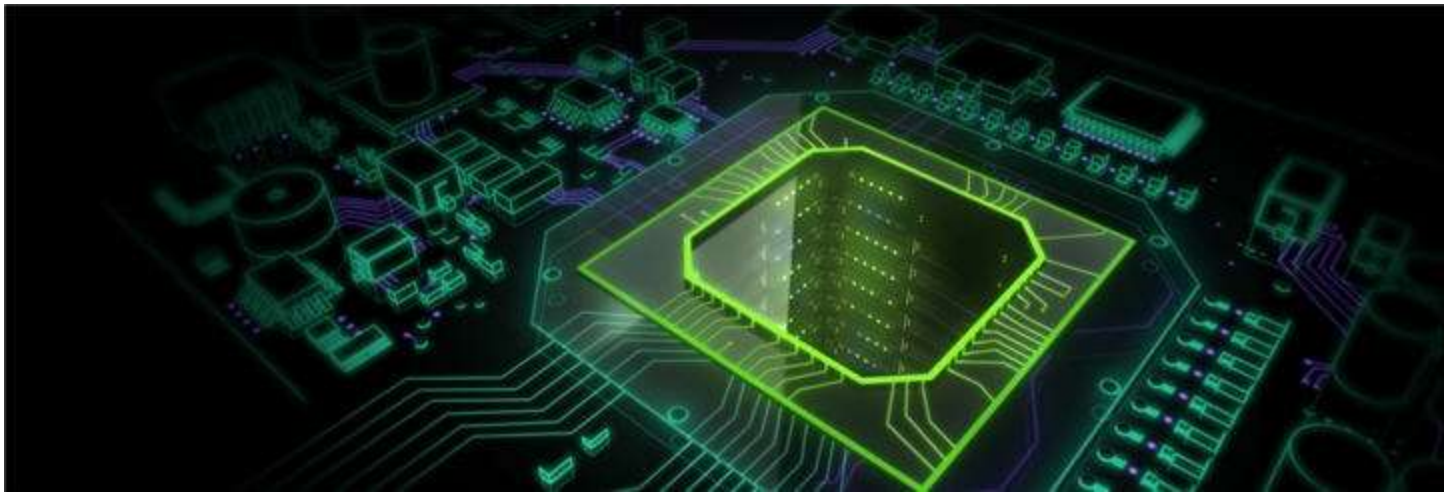
NYU
New York University

CSCI-GA.3033-012

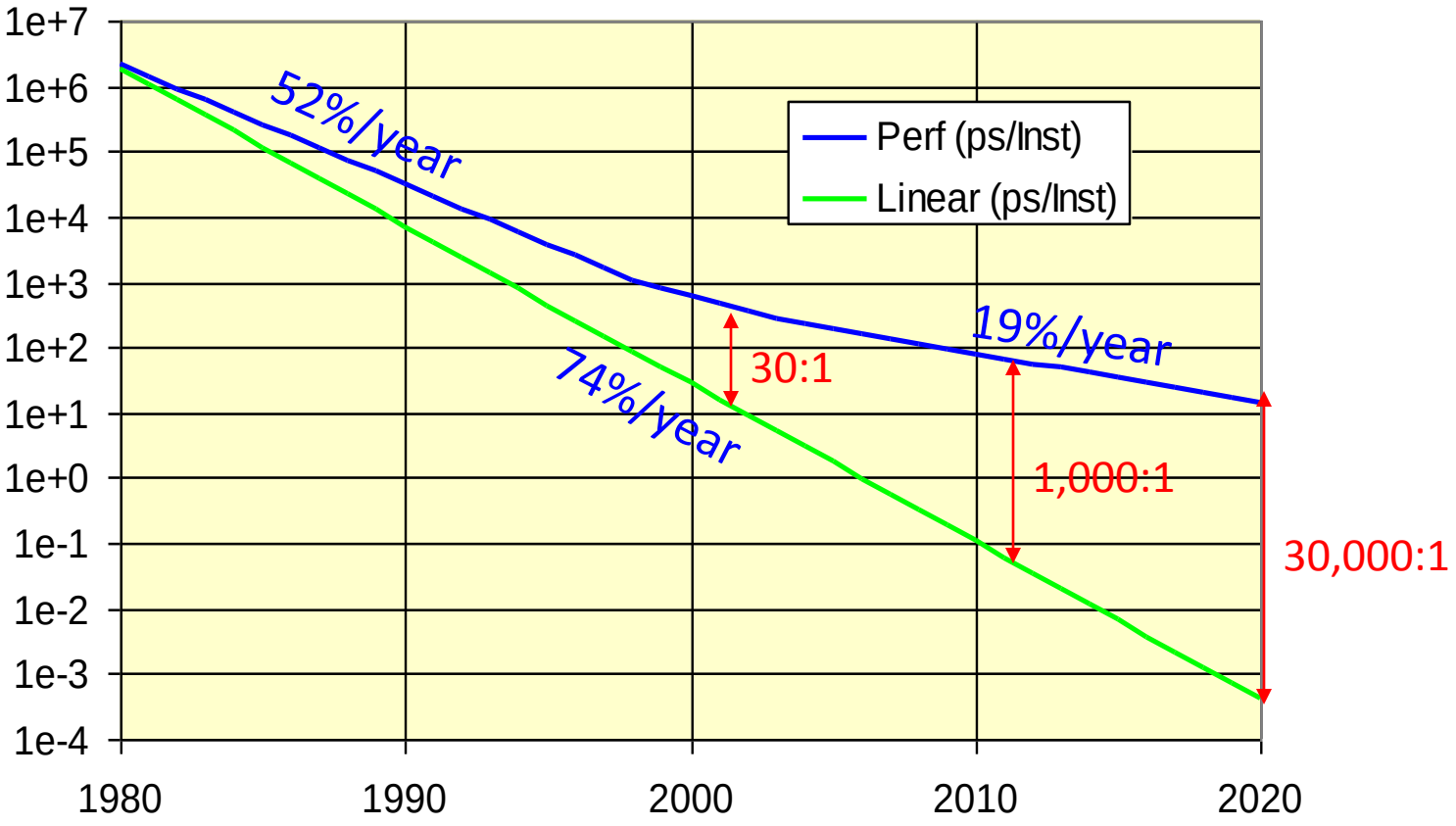
Graphics Processing Units (GPUs): Architecture and Programming

Lecture 10: Heterogeneous Systems

Mohamed Zahran (aka Z)
mzahran@cs.nyu.edu
http://www.mzahran.com

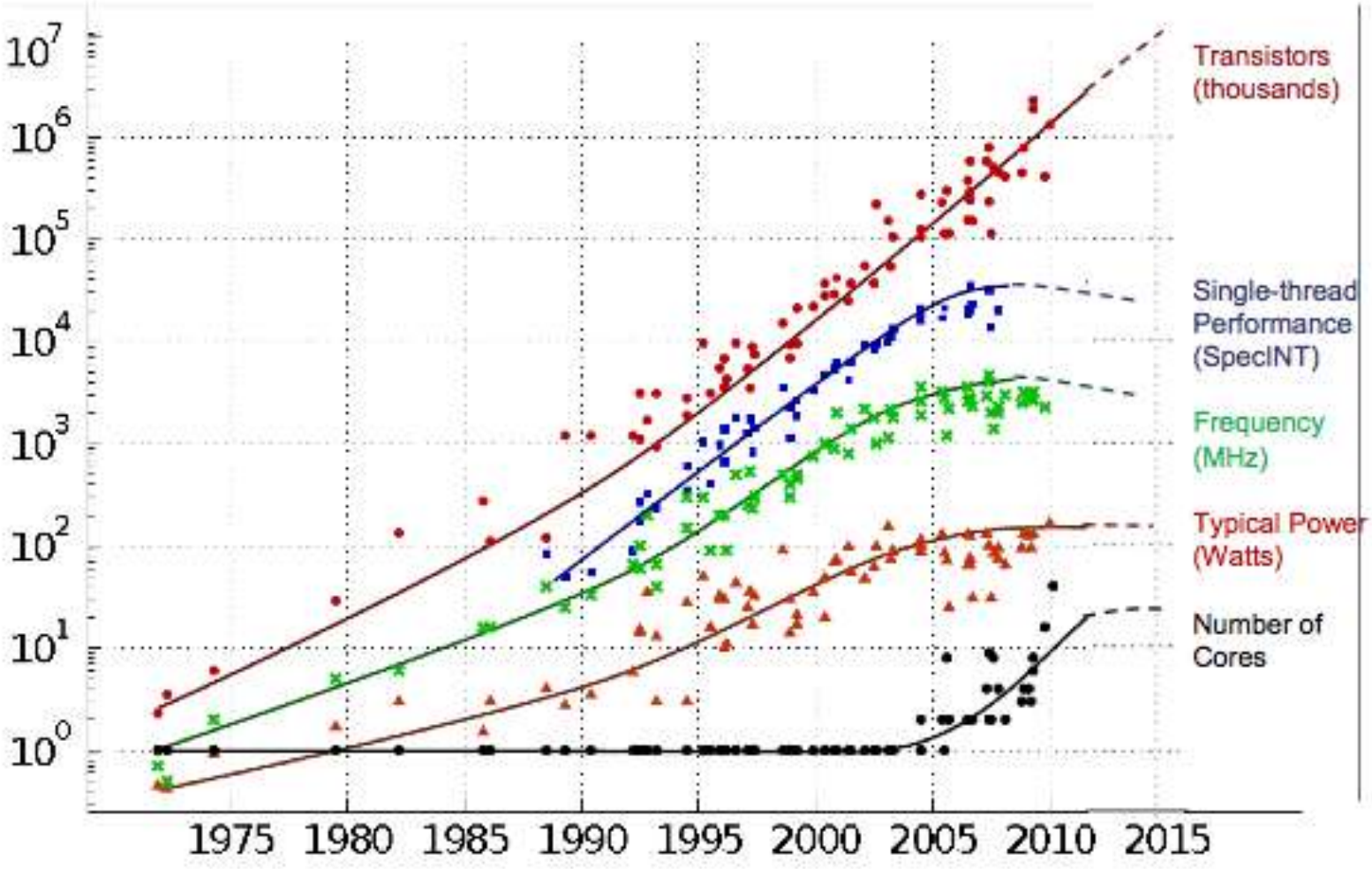


Not Enough ILP

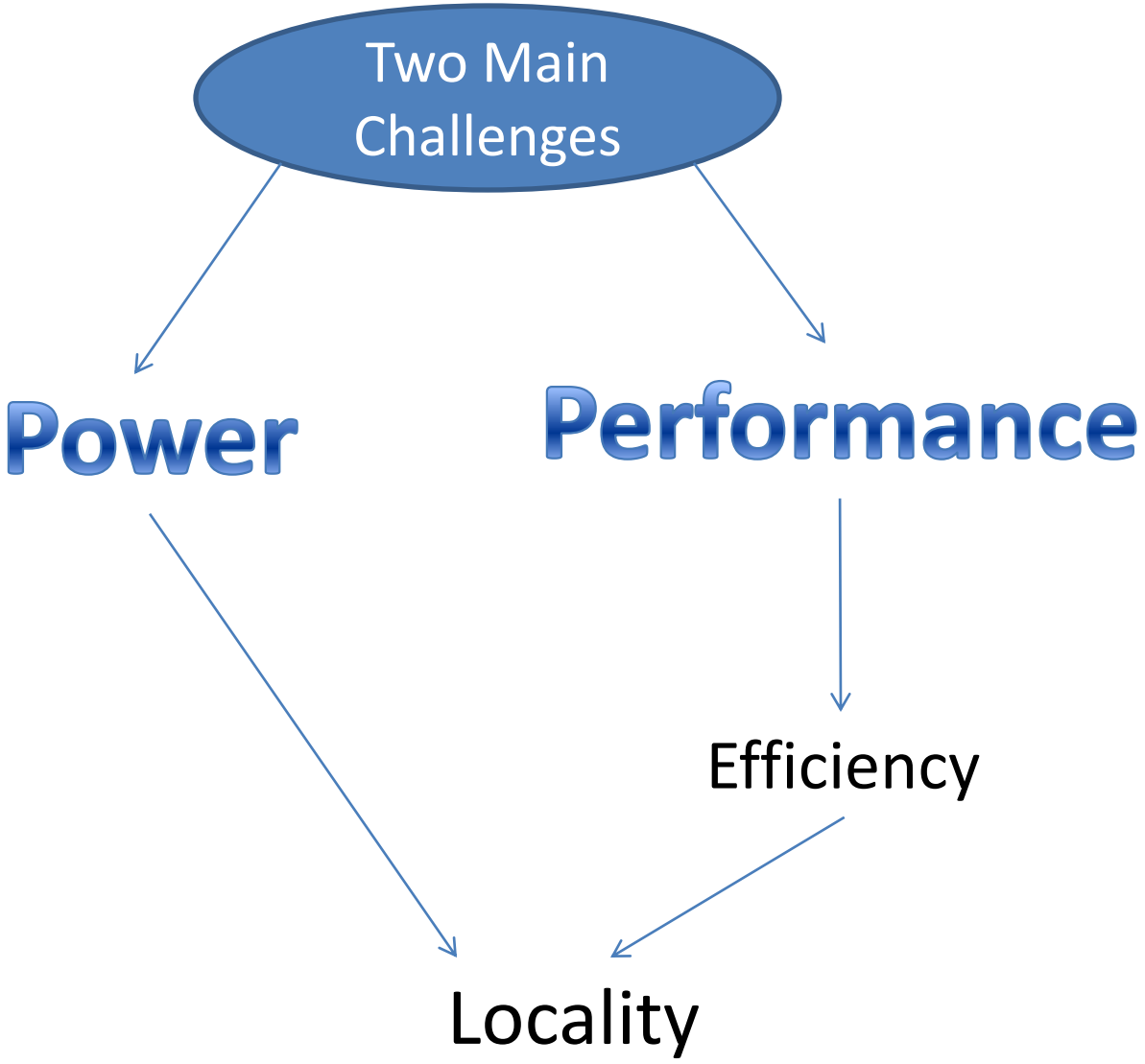


Dally et al. “The Last Classical Computer”, ISAT Study, 2001

Historical scaling has ended in 2010



Original data collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond and C. Batten
Dotted line extrapolations by C. Moore



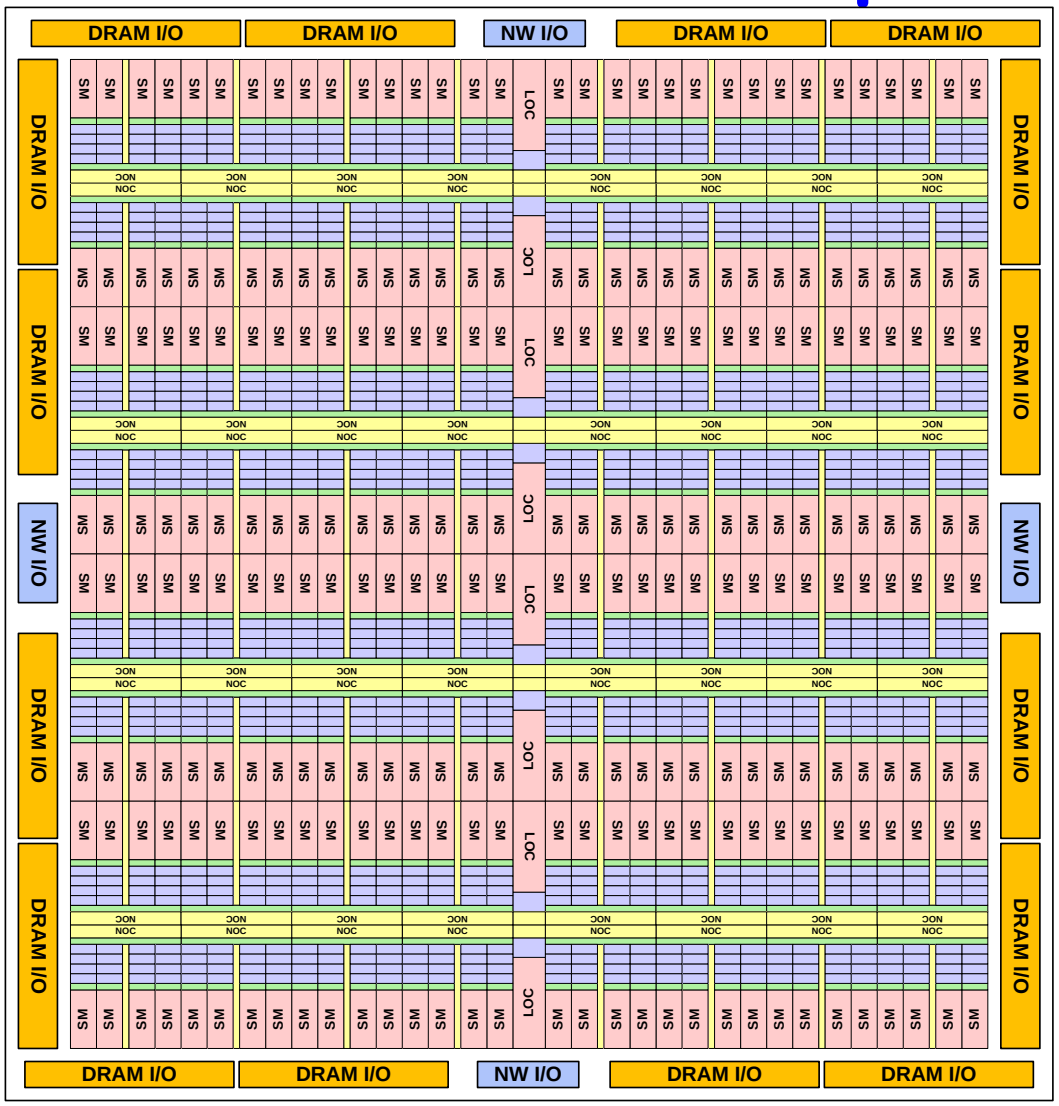
Data movement costs more than computation.

Why Heterogeneous Multicore/Manycore

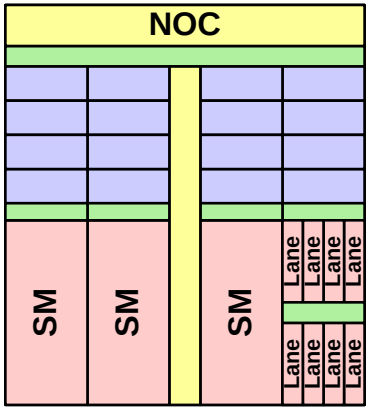
- Wide variety of software applications
- Application with fewer and sophisticated threads -> Traditional multicore with latency optimize cores
- Application with high concurrency -> large number of throughput optimized cores
- Energy efficiency
- Heterogeneous computing is needed to reach ExaScale computing

NVIDIA: ECHELON

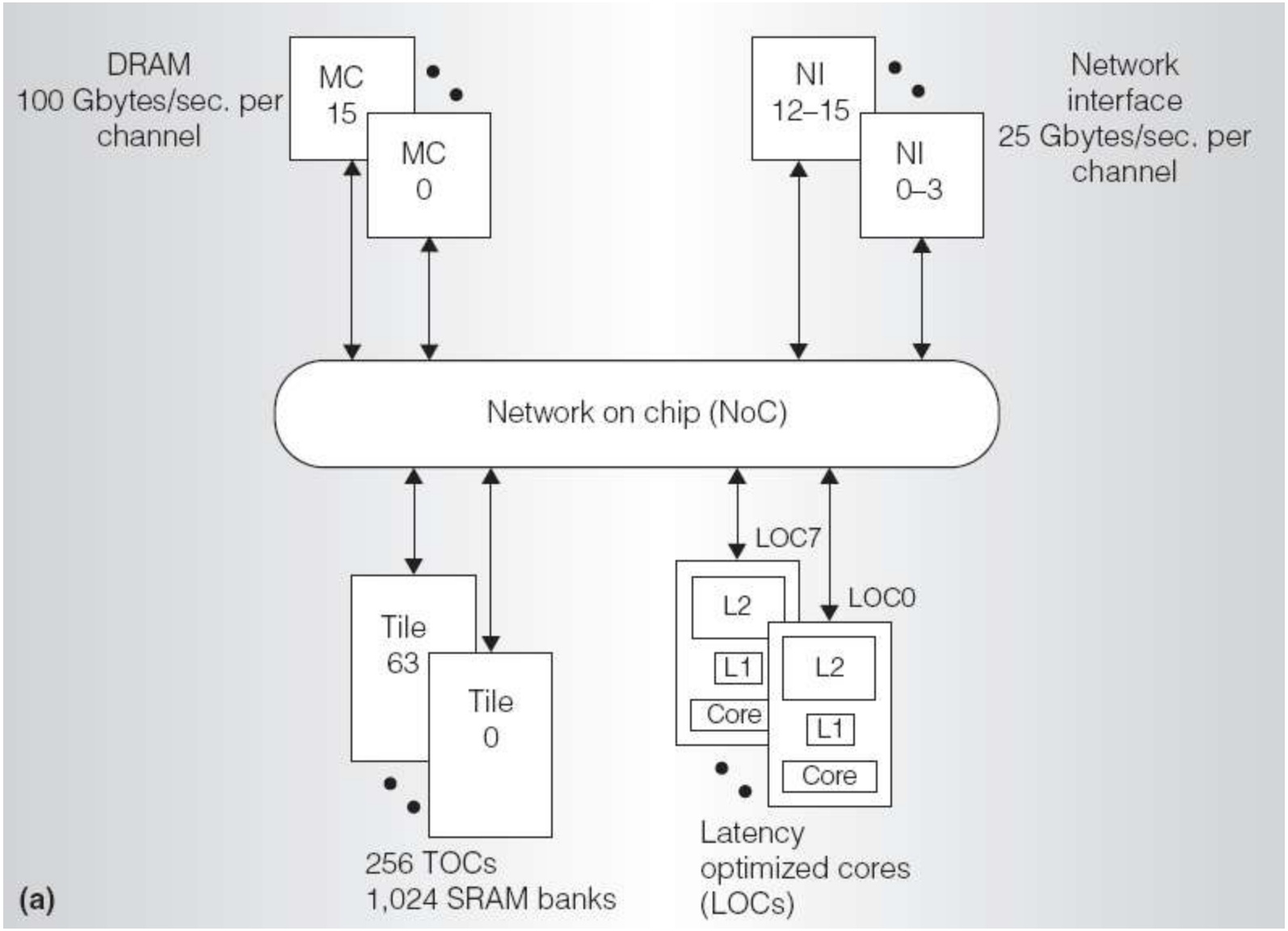
Echelon Chip Floorplan

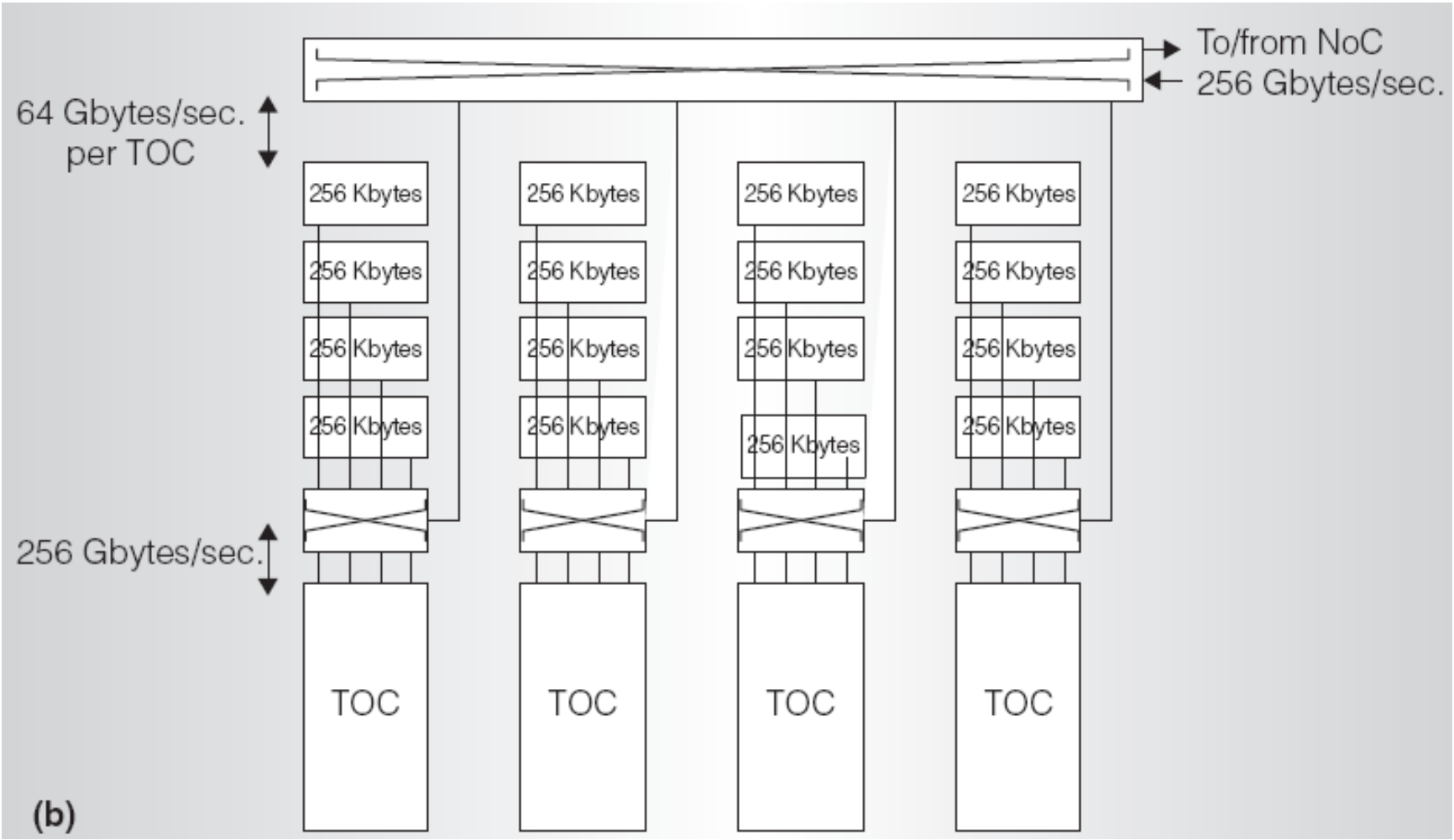


17mm



10nm process
290mm^2


DRAM
100 Gbytes/sec. per channel
MC 15
MC 0
NI 12–15
NI 0–3
Network interface
25 Gbytes/sec. per channel
Network on chip (NoC)
LOC7
LOC0
Tile 63
Tile 0
L2
L1
Core
L2
L1
Core
256 TOCs
1,024 SRAM banks
Latency optimized cores (LOCs)
(a)


To/from NoC
256 Gbytes/sec.
64 Gbytes/sec.
per TOC
256 Kbytes
256 Kbytes
256 Kbytes
256 Kbytes
256 Kbytes
256 Kbytes
256 Kbytes
256 Kbytes
256 Kbytes
256 Kbytes
256 Kbytes
256 Kbytes
256 Kbytes
256 Kbytes
256 Kbytes
256 Kbytes
256 Gbytes/sec.
TOC
TOC
TOC
TOC
(b)

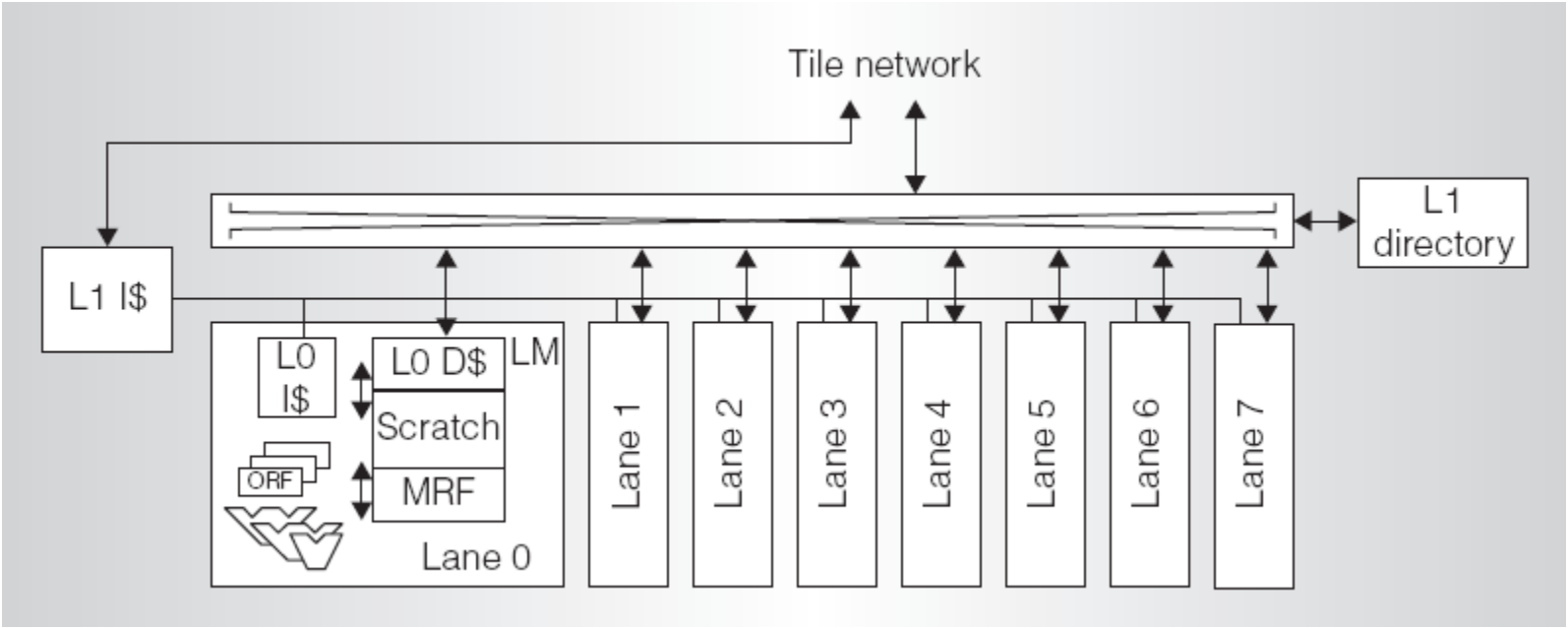
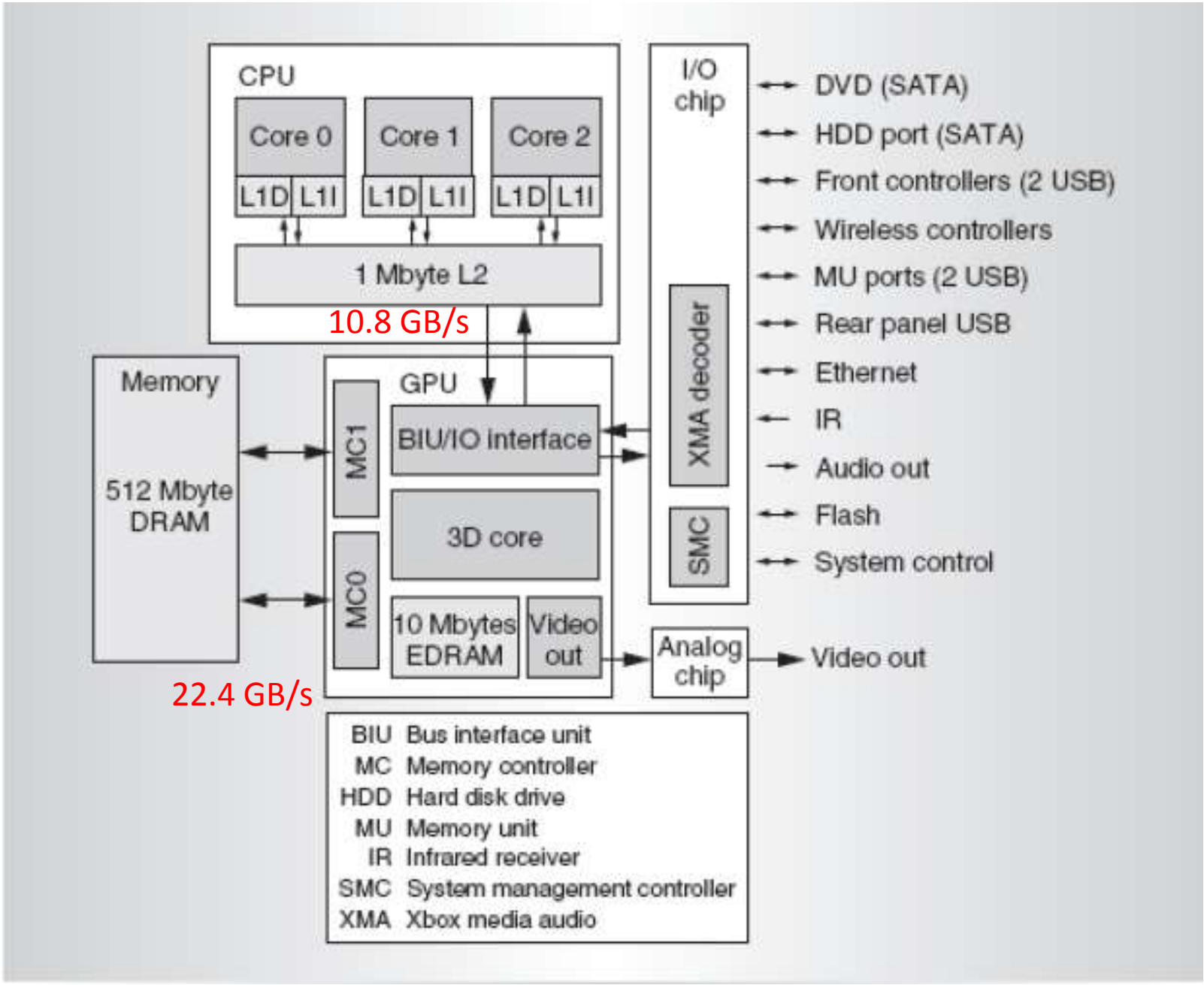

Tile network
L1 directory
L1 I$
L0 I$
L0 D$
LM
Scratch
MRF
ORF
Lane 0
Lane 1
Lane 2
Lane 3
Lane 4
Lane 5
Lane 6
Lane 7

Table 3. Echelon's parameters.

Capability	Per lane	Per TOC	Per tile	Per chip
Instructions per clock	3	24	96	6,144
Instructions per second at 2 GHz	6	48	192	12,288
Double-precision floating-point operations per clock	4	32	128	8,192
Gflops at 2 GHz	8	64	256	16,384
Total threads	64	512	2,048	131,072
Active threads	4	32	128	8,192
Data SRAM (Kbytes)	32	256	5,120	327,680
L0 instruction cache (Kbytes)	1	8	32	2,048
L1 instruction cache (Kbytes)	—	32	128	8,192
DRAM bandwidth (Gbytes/second)	—	—	—	1,600
Network interface bandwidth (Gbytes/second)	—	—	—	400

XBOX 360


CPU
Core 0
Core 1
Core 2
L1D
L1I
1 Mbyte L2
10.8 GB/s
Memory
512 Mbyte DRAM
22.4 GB/s
GPU
MC1
MC0
BIU/IO interface
3D core
10 Mbytes EDRAM
Video out
I/O chip
XMA decoder
SMC
Analog chip
DVD (SATA)
HDD port (SATA)
Front controllers (2 USB)
Wireless controllers
MU ports (2 USB)
Rear panel USB
Ethernet
IR
Audio out
Flash
System control
Video out
BIU Bus interface unit
MC Memory controller
HDD Hard disk drive
MU Memory unit
IR Infrared receiver
SMC System management controller
XMA Xbox media audio

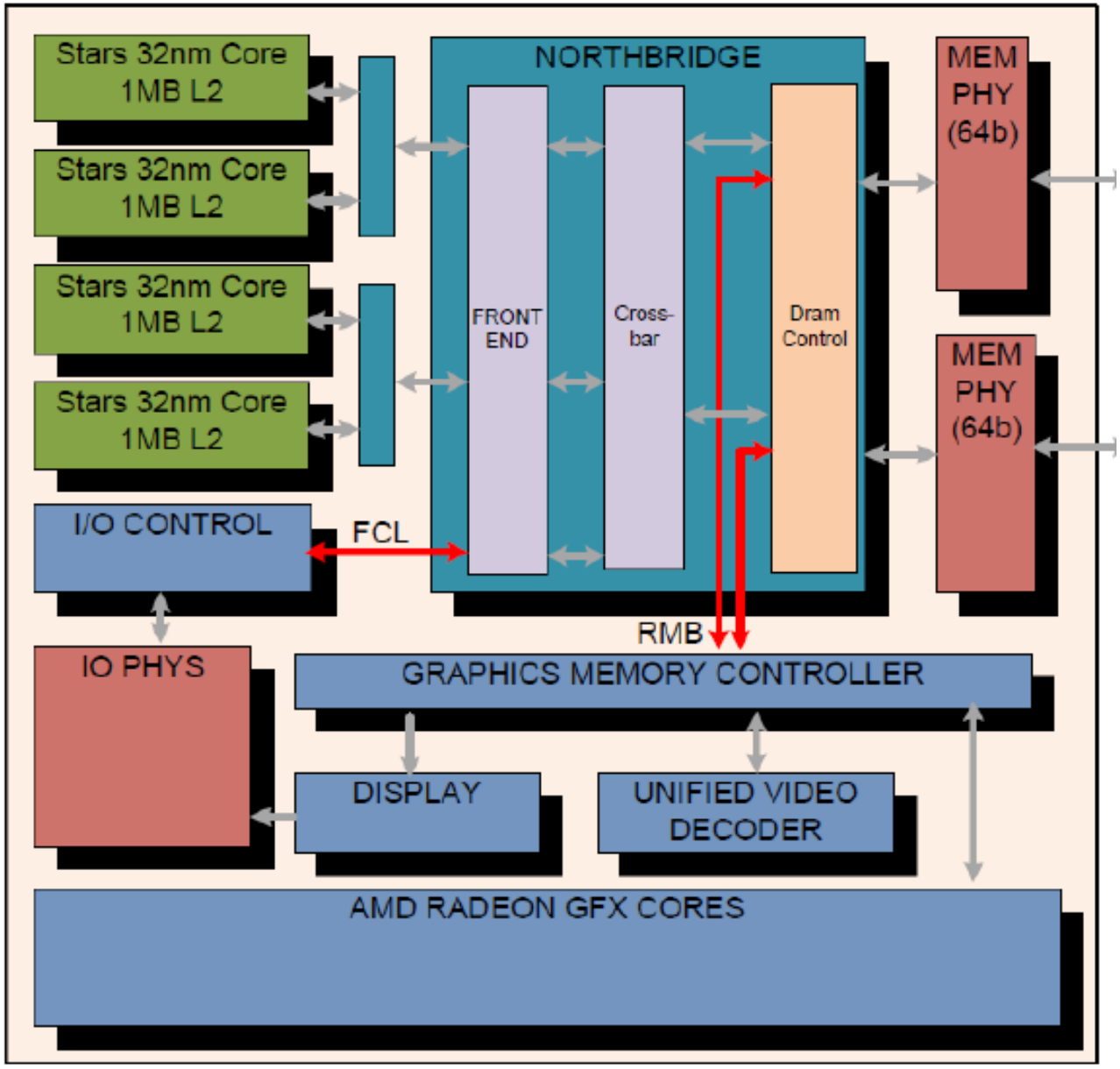
Specifications

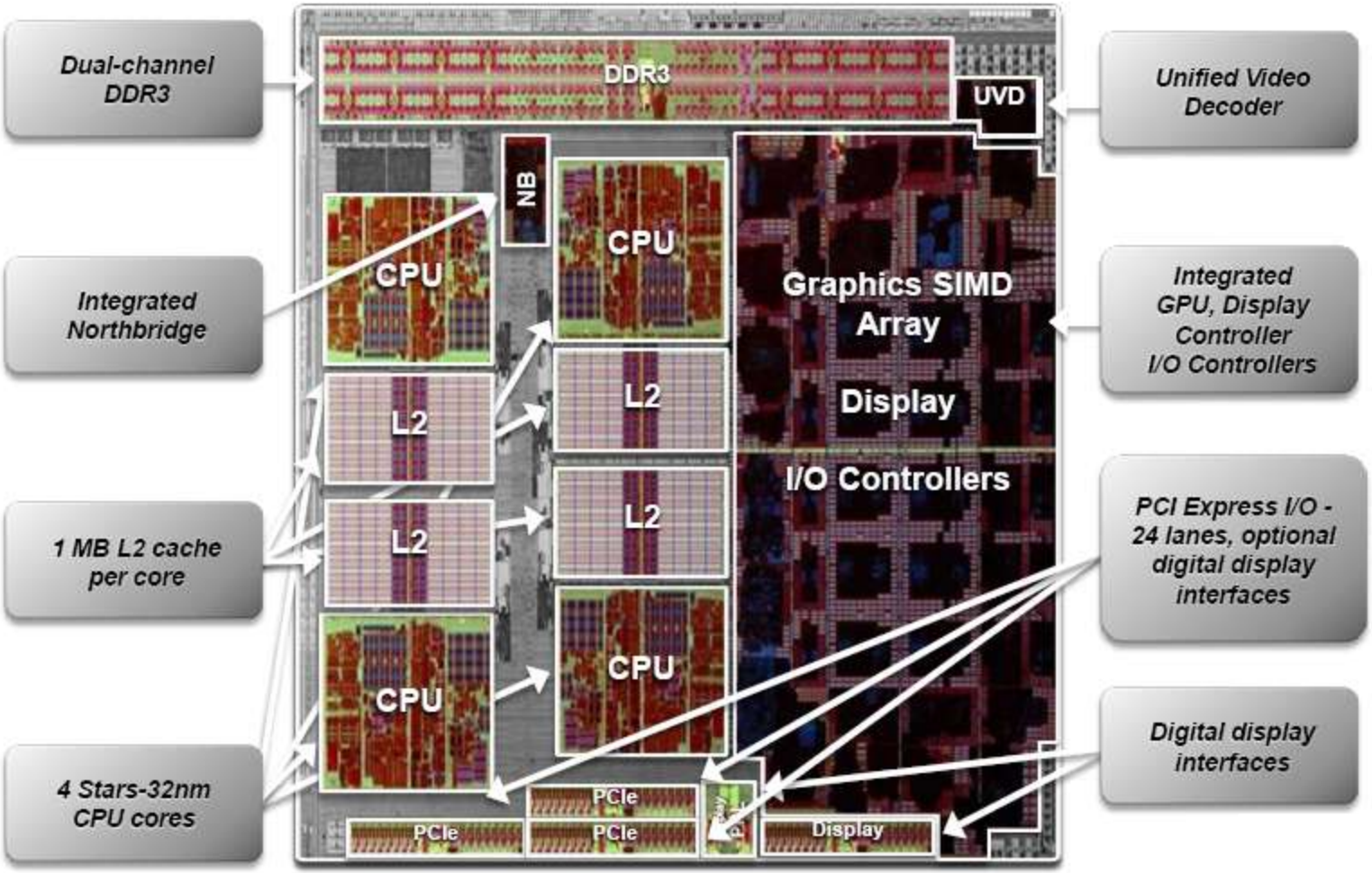
- 3 CPU cores
  - 4-way SIMD vector units
  - 8-way 1MB L2 cache (3.2 GHz)
  - 2 way SMT
  - In-order
  - 2 Instructions/cycle
- ATI GPU with embedded EDRAM
- 3D graphics units
- 512-Mbyte DRAM main memory

Philosophy

- Value for 5-7 years
- Big performance increase over last generation
- Support high-definition video
- extremely high pixel fill rate (goal: 100+ million pixels/s)
- Flexible to suit dynamic range of games
- balance hardware, homogenous resources
- Programmability (easy to program)

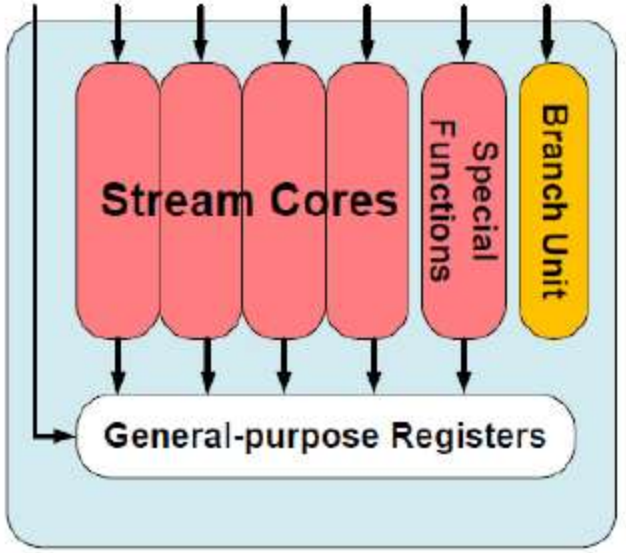
AMD'S "LLANO" FUSION APU


Stars 32nm Core
1MB L2
Stars 32nm Core
1MB L2
Stars 32nm Core
1MB L2
Stars 32nm Core
1MB L2
NORTHBRIDGE
FRONT END
Cross-bar
Dram Control
MEM PHY (64b)
MEM PHY (64b)
I/O CONTROL
FCL
RMB
IO PHYS
GRAPHICS MEMORY CONTROLLER
DISPLAY
UNIFIED VIDEO DECODER
AMD RADEON GFX CORES


Dual-channel DDR3
Integrated Northbridge
1 MB L2 cache per core
4 Stars-32nm CPU cores
DDR3
UVD
NB
CPU
CPU
CPU
CPU
L2
L2
L2
Graphics SIMD Array
Display
I/O Controllers
PCIe
PCIe
PCIe
Display
Unified Video Decoder
Integrated GPU, Display Controller I/O Controllers
PCI Express I/O - 24 lanes, optional digital display interfaces
Digital display interfaces

AMD – Radeon GPU VLIW-5

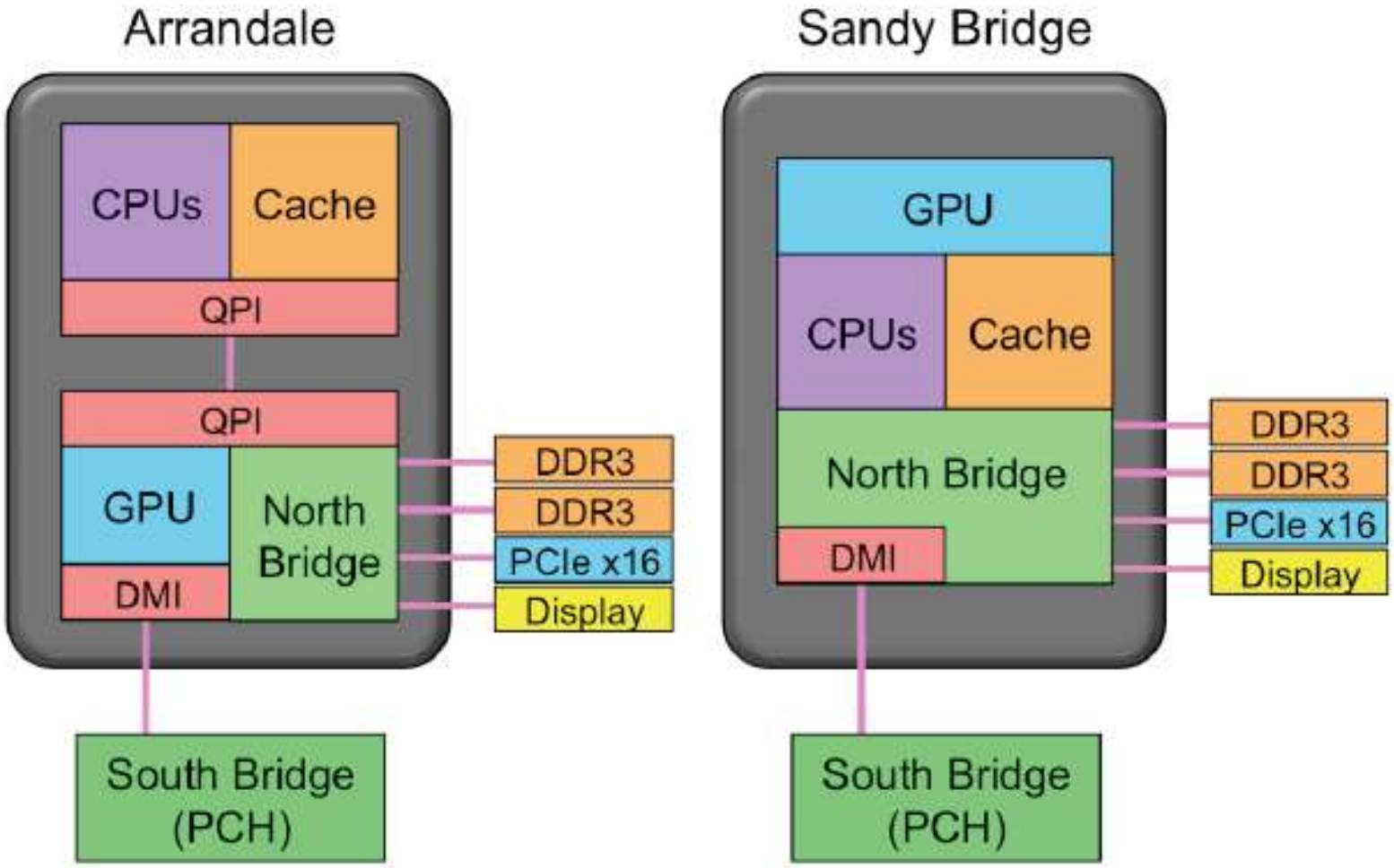
- Includes 4 Stream Cores
  - 1 Special Functions Stream Core
  - Branch Unit
  - General Purpose Registers
- 4 Stream Cores are capable of
  - 4 32-bit FP MULADD per clock
  - 4 24-bit IntMUL or ADD per clock
  - 2 64-bit FP MUL or ADD per clock
  - 1 64-bit FP MULADD per clock
- Additional special function core 32b-FP MULADD per clock



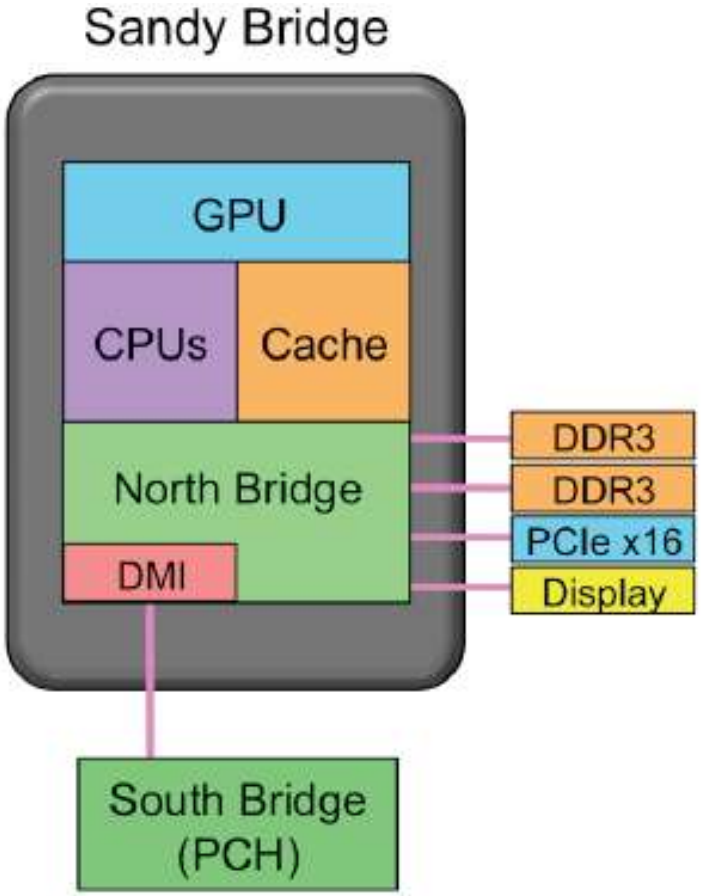
INTEL: SANDY BRIDGE

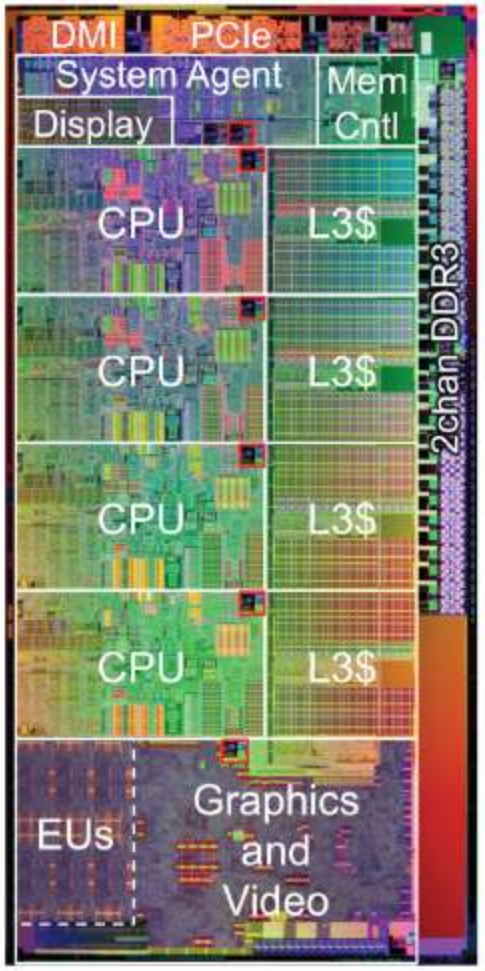
Sandy Bridge

- Intel Tock
  - Intel's first to GPU on the processor itself.
- Improvement over its predecessor Nehalem
- Targeting multimedia applications
  - Introduced Advanced Vector Extensions (AVX)
- More power-efficient than Westmere


Arrandale
CPUs
Cache
QPI
QPI
GPU
North Bridge
DMI
DDR3
DDR3
PCIe x16
Display
South Bridge (PCH)
Sandy Bridge
GPU
CPUs
Cache
North Bridge
DMI
DDR3
DDR3
PCIe x16
Display
South Bridge (PCH)

- The GPU can access the large L3 cache
- Intel's team totally re-designed the GPU





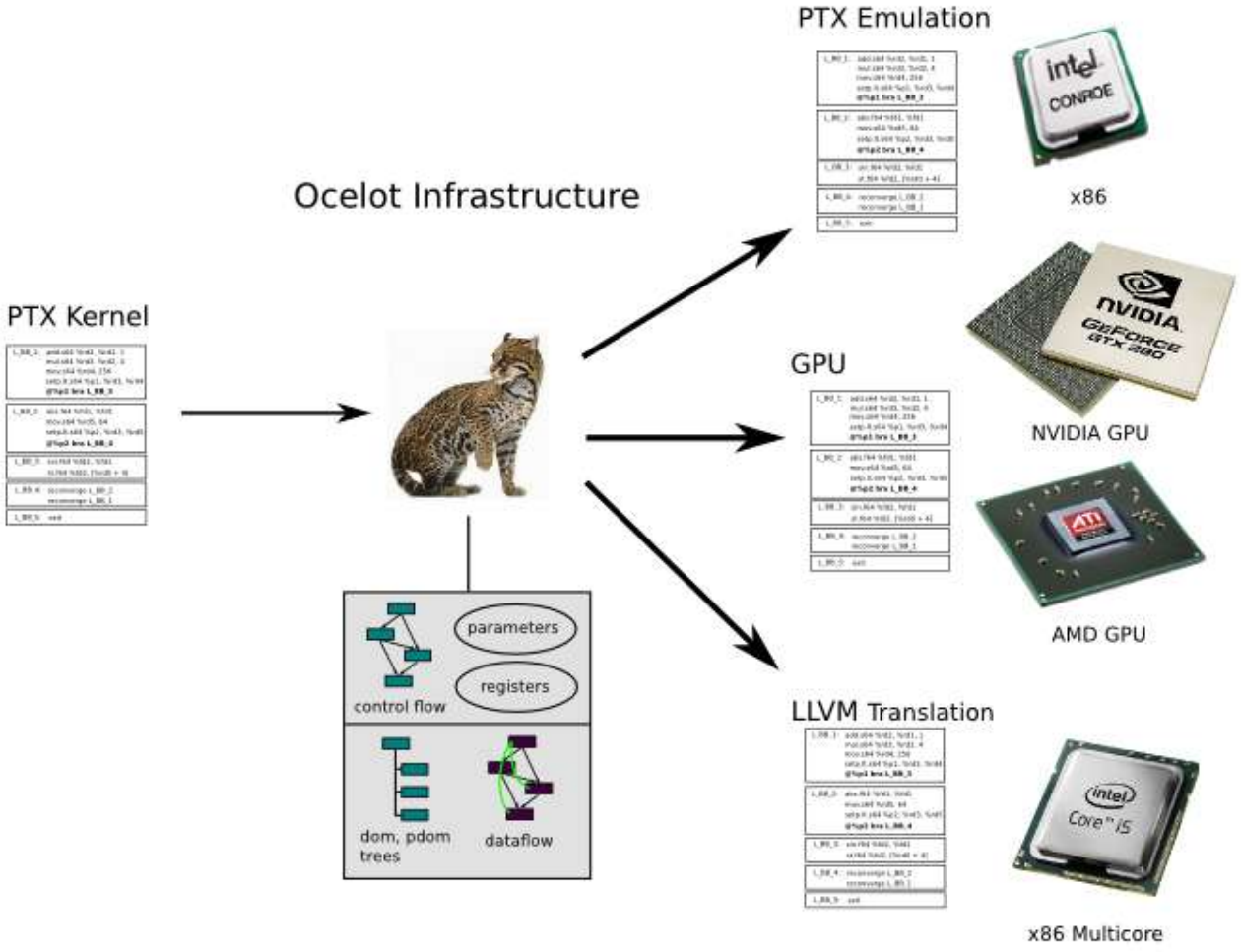
What is/are the main design philosophy/philosophies of all those heterogeneous systems?

Which programs benefit the most from them?

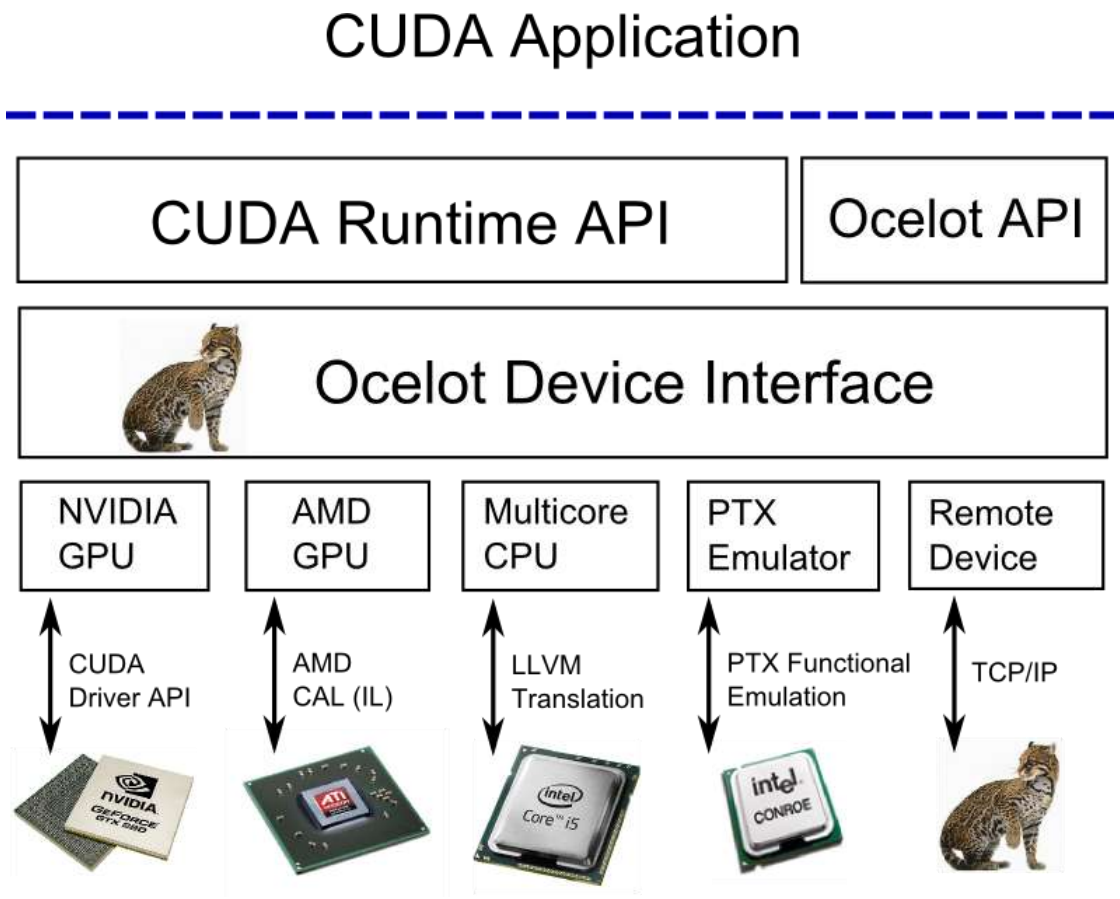
Programming Heterogeneous Systems

- Who's doing what?
  - programmer
  - OS
  - Compiler
- SSE/AVX should be appropriate if the process is suitable for SIMD parallelization and not for GPU
- Several choices:
  - OpenMP
  - pthreads
  - CUDA
  - OpenCL
  - Intel threading Building Block (TBB)

Useful Tool: gpuocelot



Useful Tool: gpuocelot



Conclusions

- It is all about performance and power
- Algorithms should be designed to perform more work per unit data movement.
  - Data movement dominates power.
- Issues:
  - How do GPU and CPU share memory?
  - Bandwidth (on-chip and off-chip)
  - Coherence?
  - Portability of code?