INF5063:
Memory Considerations

Håvard Espeland

August 28, 2012

Volatile Memory Types

- SRAM
  - Low latency, high throughput, runs at clock speed.
  - Typically used in caches.
  - Usually small (KB-MB) and expensive (4-6 transistors/bit).
- DRAM
  - High latency, low throughput, runs asynchronously.
  - Large system memory (MB-TB).
  - Cheap (1 capacitor and 1 transistor / bit).

Memory Transfers

- Direct addressing (PIO)
  - Request data from memory using an address or register indirect.
- Directed Memory Access (DMA)
  - Transfer directly, without involving the CPU.
- Prefetching
  - Copy data to cache before it is used. Can be done implicitly (hardware prefetch) or explicitly (prefetch instructions).
- Cache writeback
  - Write back to memory when cache object is evicted.

Prefetching

- Hide latency by copying the data to cache before you actually need it.
- Implicit prefetching works really well on modern x86.
  - You can do it explicitly, but can be hard / impossible to beat hardware prefetcher.
- Not always so on other processors.
  - E.g. PowerPC (Xbox360, PS3, Mars Curiosity) often greatly benefit from explicit prefetching.

Cache Coherency

- Ensures that changes in one cache are reflected for all others.
  - Expensive!
  - Can be done in software or hardware.
  - Typically done by broadcasting writes to other caches.
- Easy to use for programmers
  - But hard to scale...
- Often available on SMP systems such as x86.
  - But not always.

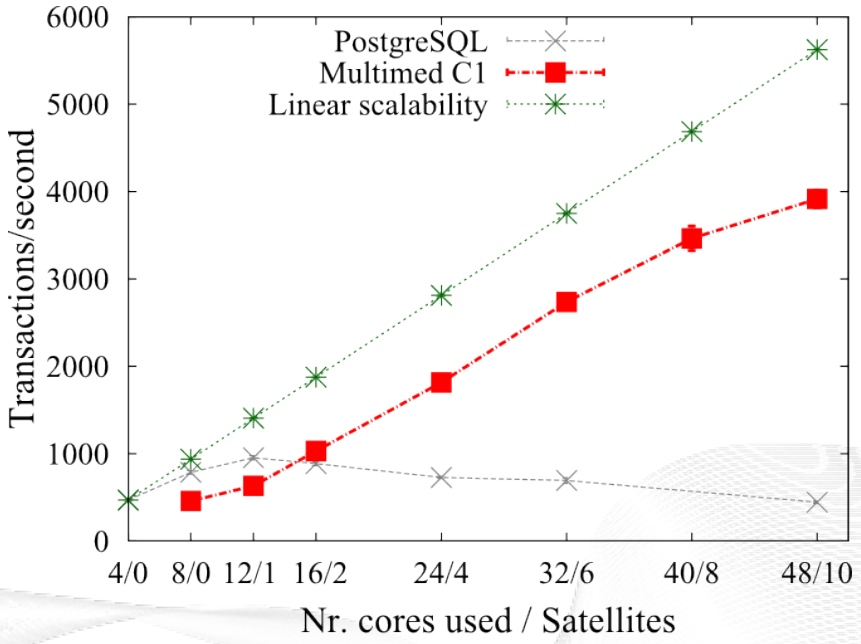
Non-Uniform Memory Access

- Memory acccess time depends on location relative to processor.
  - Memory can be off on-chip, off-chip or even on a remote node.
  - Widely used in HPC.
- Modern multiprocessors incorporate NUMA concepts.
  - Several types of memory on chip, e.g. Cell, GPU, IXP, etc.
- Remote cache coherent NUMA on standard x86 is possible
  - NUMAScale exposes remote CPUs and RAM by tapping the HyperTransport bus directly.


ESSO
ESSO
A

Cache Coherent Scalability

- As cache contention grows, scalability suffers.
  - Database performance on 12-core AMD Magny Cours.
  - Shows PostgreSQL, similar results with MySQL.
  - Multimed replicates databases and runs independent instances.



(a) Scalability throughput: 2GB, 200 clients

Database Engines on Multicores, Why Parallelize When You Can Distribute?, Salomie et al., Eurosys 2011.

Architectures Without Coherency

- Asymmetric Processors
  - Cell Broadband Engine and GPUs
- Network Processors
  - E.g., Netronome IXP
- x86 Processors
  - Intel's Single Chip Cloud Computer (SCC) with 48 cores.
- Embedded systems, DSPs and stream processors.

Transactional Memory

- Group several memory operations together atomically.
  - Replaces mutexes.
- Added to C++11
  - __transaction_atomic { c = a - b; }
- Hardware support on Intel's Haswell microarchitecture, eta:

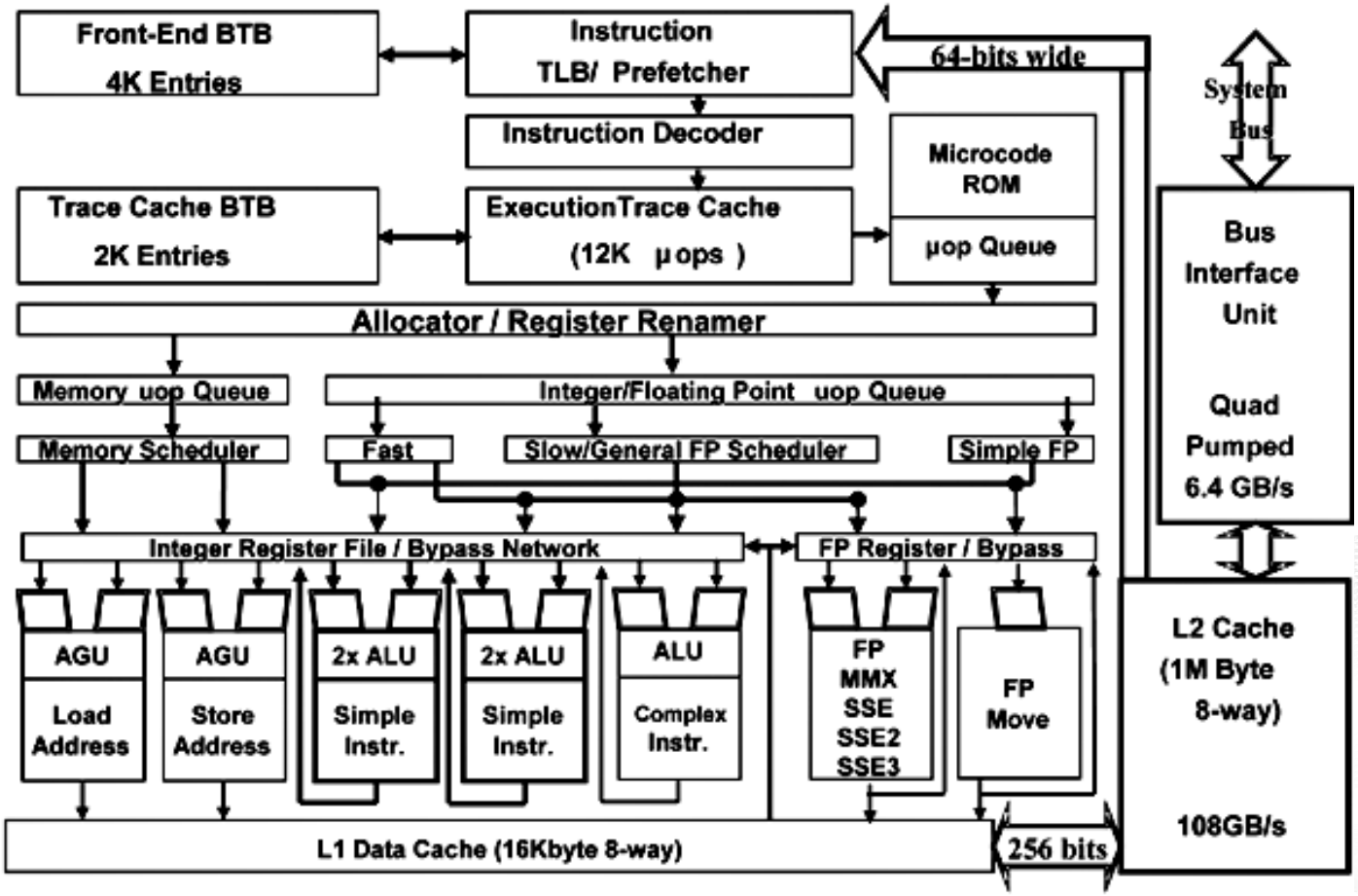
  Q2 2013.

Performance Evaluation

- Use Hardware Performance Counters
  - Libpfm exposes counters to user space without requiring a kernel module / root access.
  - A lot of registers to track.
- Tools available
  - Intel Performance Counter Monitor
  - Intel Parallel Studio

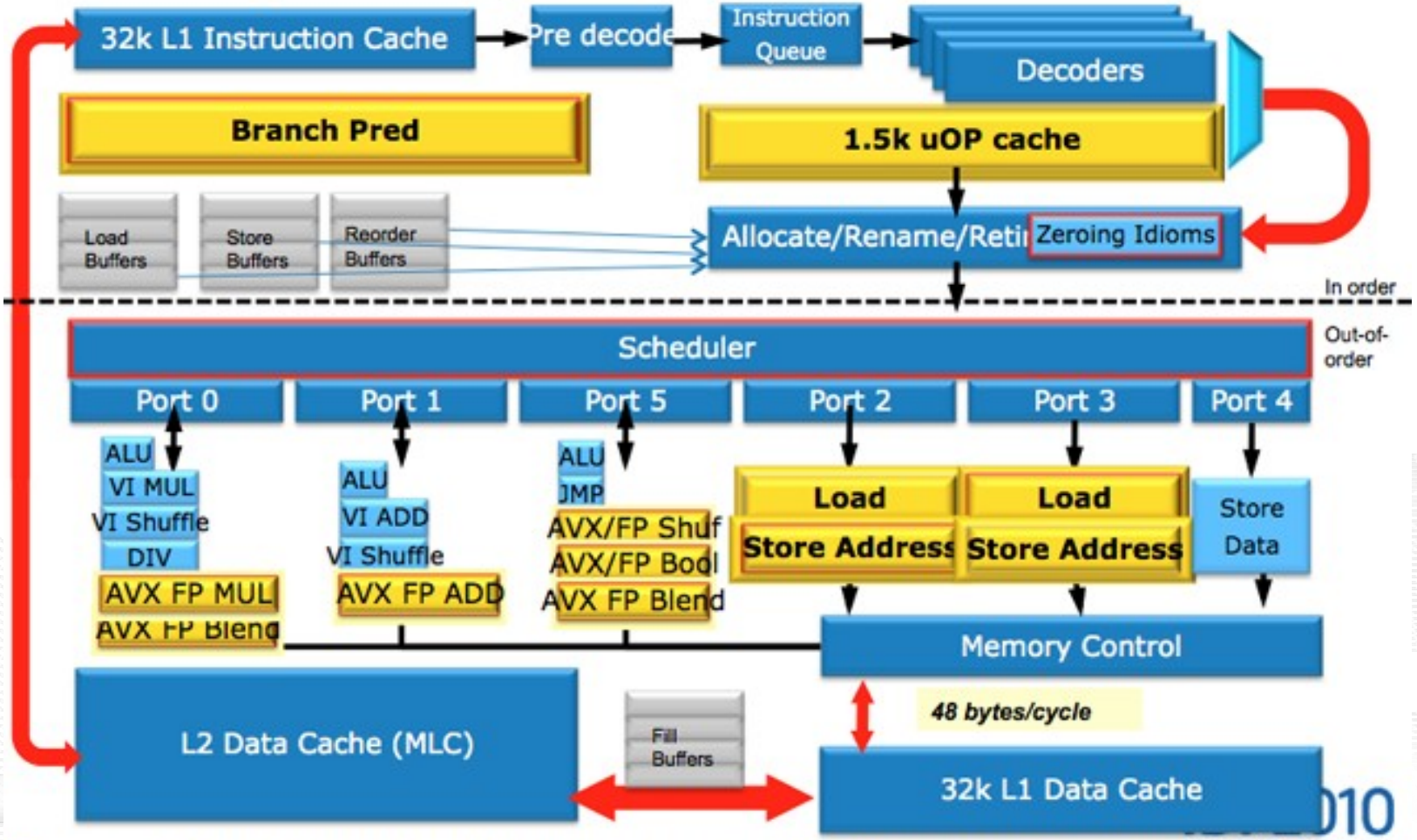
```
More info: Performance Analysis Guide for Intel® CoreTM i7
Processor and Intel® XeonTM 5500 processors
```

CPU Cache

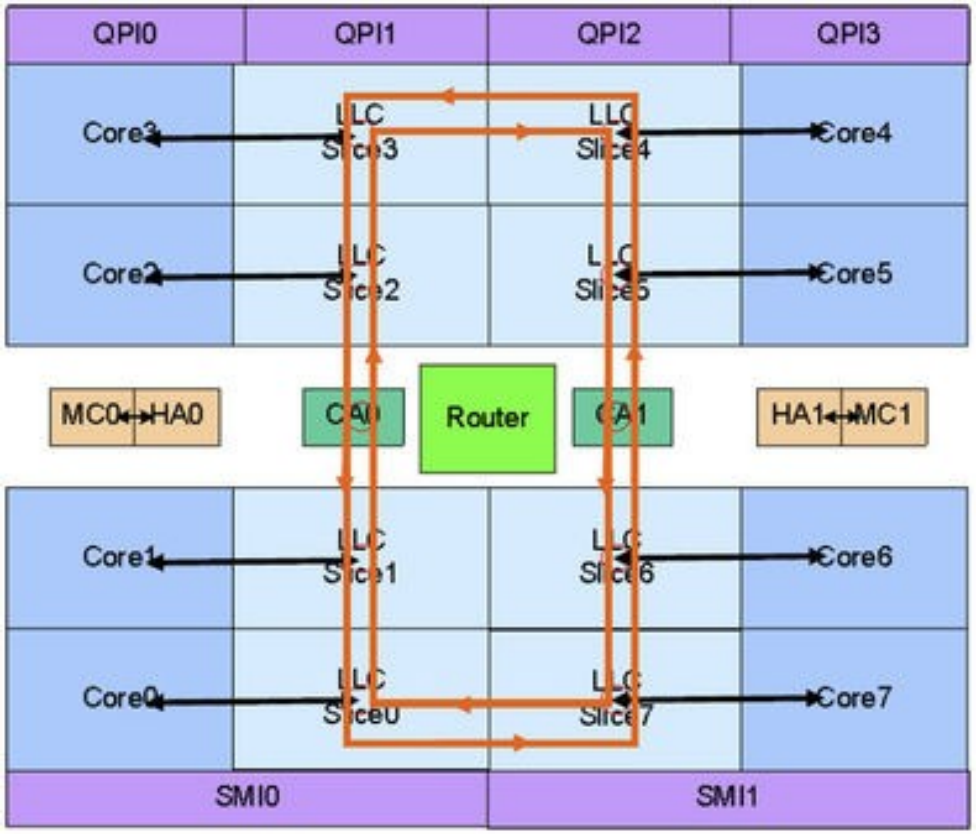


Pentium 4 Prescott

Putting it together
Sandy Bridge Microarchitecture

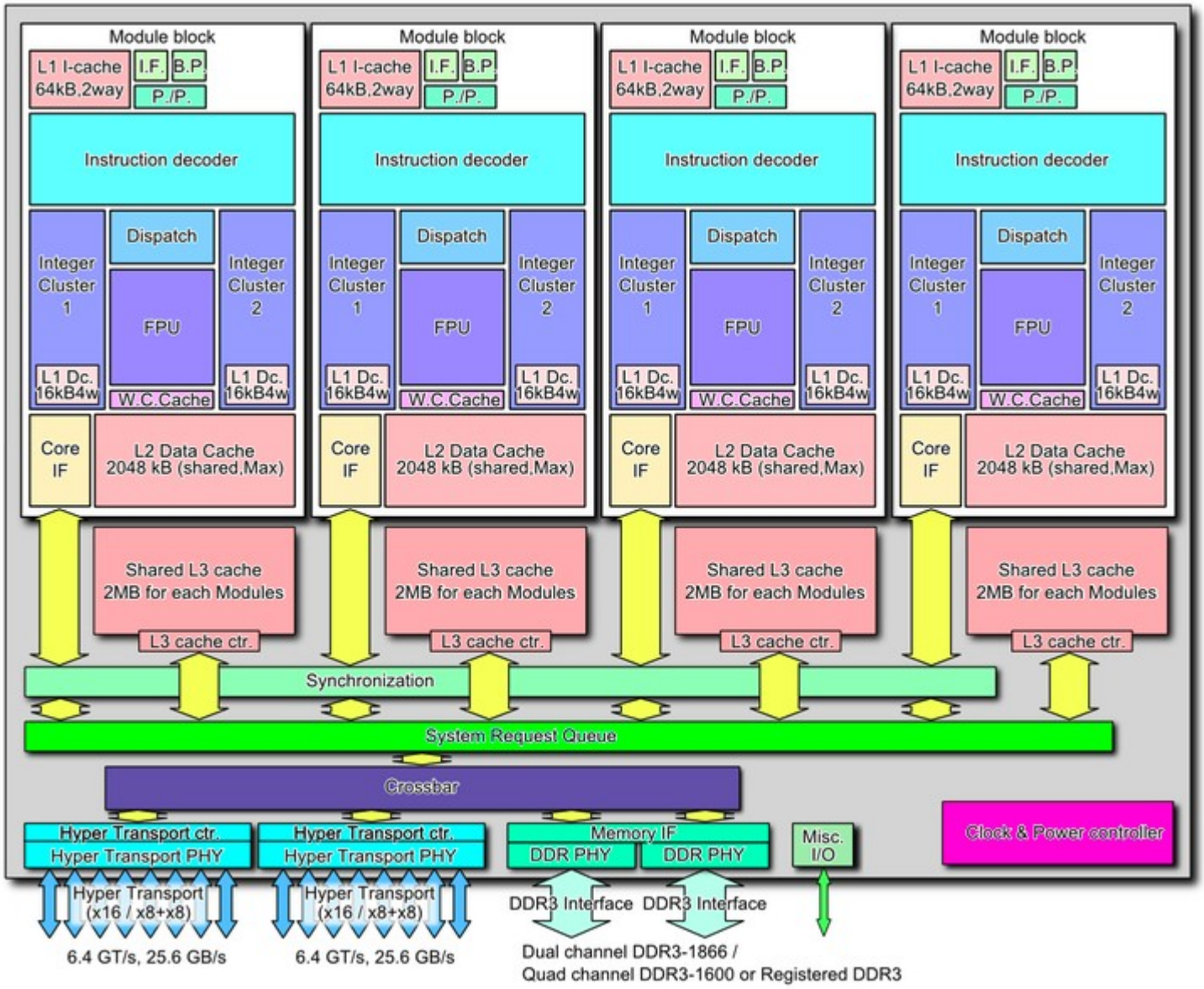


Sandy Bridge (cont)

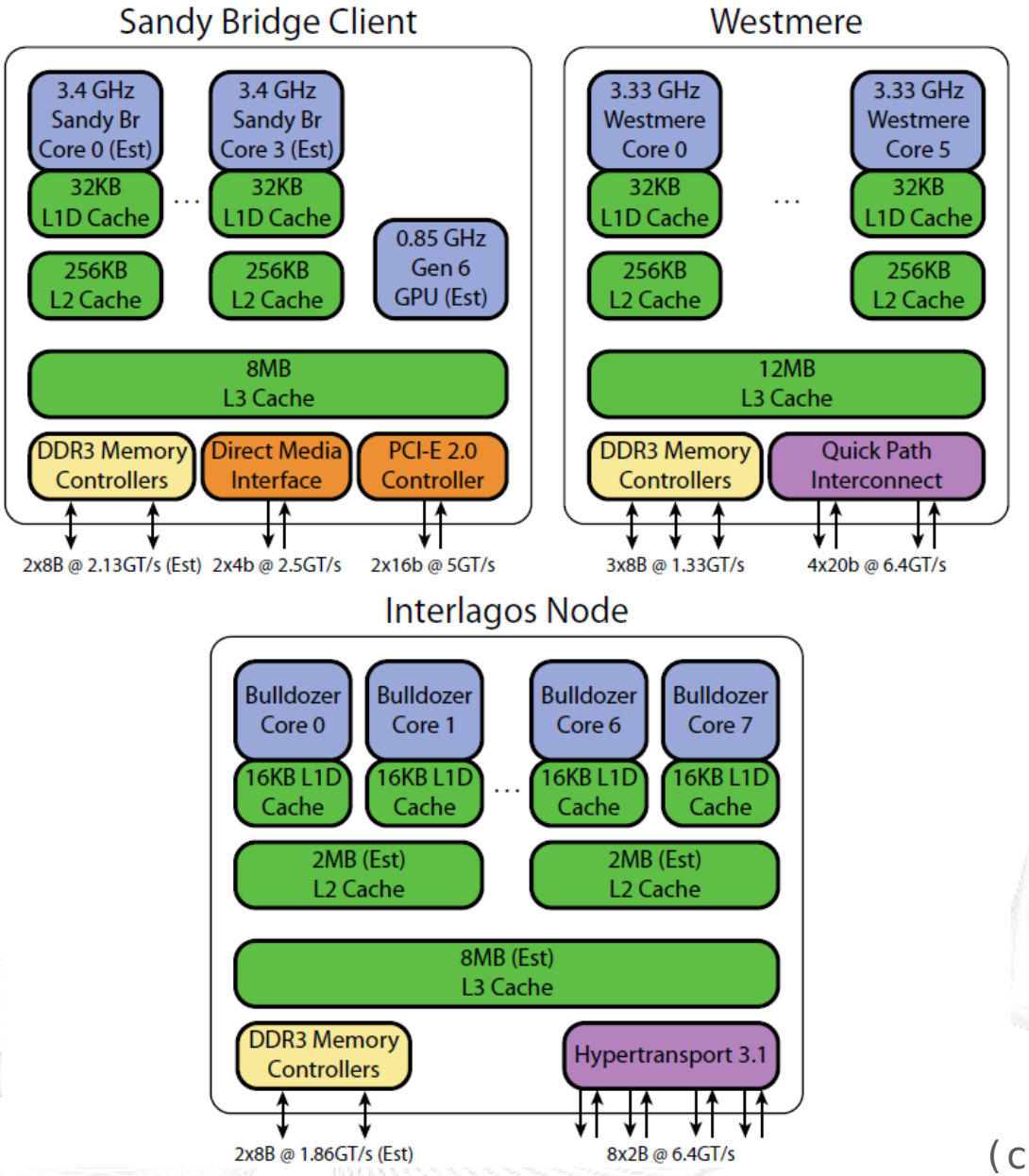


Sandy Bridge Cache latency

Size	Latency	Description
32 K	4 | 5	TLB + L1
256 K	12	+7 (L2)
2 M	36	+17 (L3) +7 (L1 TLB miss)
6 M	46	+10 (L2 TLB miss)
64 M	46 + 65 ns	+ 65 ns (RAM)
...	58 + 65 ns	+ 12 (PDE cache miss)

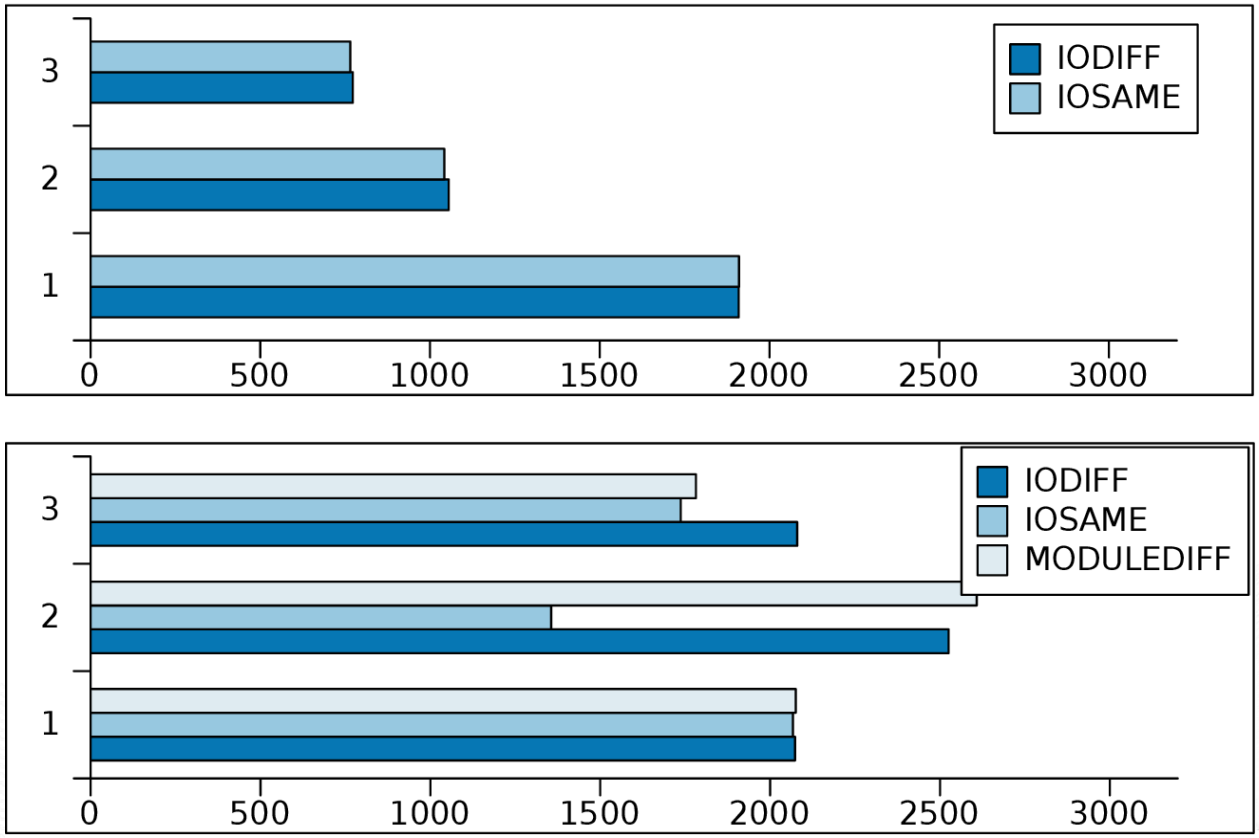


CC by Shigeru23



(c) Extremetech

Heterogeneity on x86



Placements of threads on Sandy Bridge and Bulldozer, to appear at SRMPDS12.


UNLOCKED
FX
PROCESSOR
AMD

Conclusion

- We live in a heterogeneous world.
  - Multiprocessors are becoming mainstream.
  - Even x86 behaviour varies with microarchitectures.
- Understanding the architecture is key to performance.
  - Performance is not always paramount, and you may end up using a high-level language such as python, java or whatever.
  - When resources are limited, priorities quickly change.