Supercomputer Scaling

Life in the Fast Lane

Cluster Technology

Supercomputers for the masses:
Get a bunch of cheap PCs
String a lot of cable
Voila! A supercomputer!

Cluster Technology

Supercomputers for the masses:
Get a bunch of cheap PCs
String a lot of cable
Voila! A supercomputer!

Or maybe not…

Cluster Technology

Failures multiply
Always a few dead nodes
Can turn into many

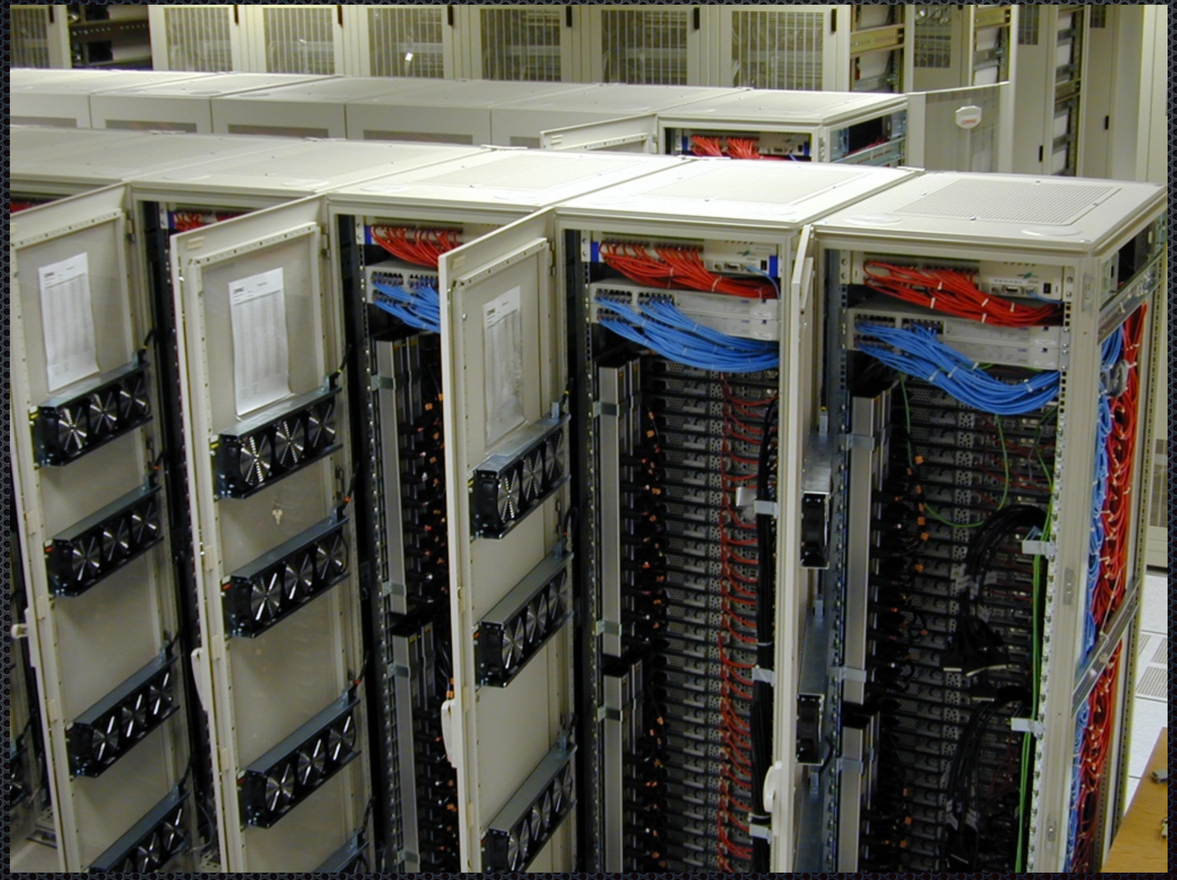
Considerations:
Heat removal
Power conditioning
Power sequencing
Vibration

Cluster Technology

Communication
Ethernet is high latency
Infiniband

Cheap PCs often use
multiple chipsets

A Rack Cluster



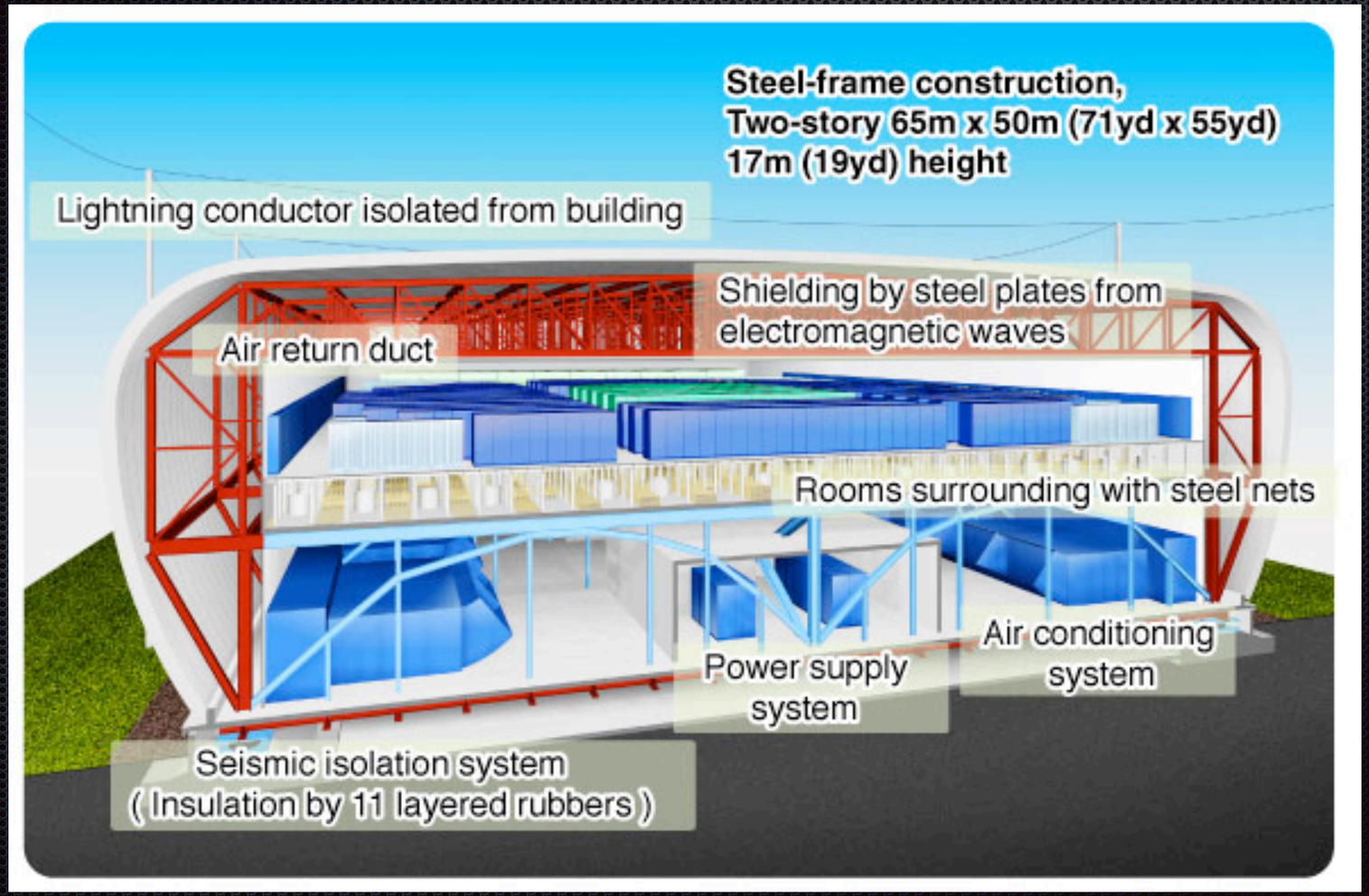
A Computer Room

A Computer

A Computer Power Supply

Another Computer

A Computer Building



Earth Simulator 2

NEC SX-9 vector processors

1280 Nodes
131 TFlops
8 CPUs/node
4-way
superscalar
256-way vector

Original Teraflop

ASCI Red: ~9K Pentium II, 3D mesh

ASC-Q

ASCI Purple

~12K Power5, 100 TFLOPS, 50 TB mem, 2 PB disk, 7.5 MW

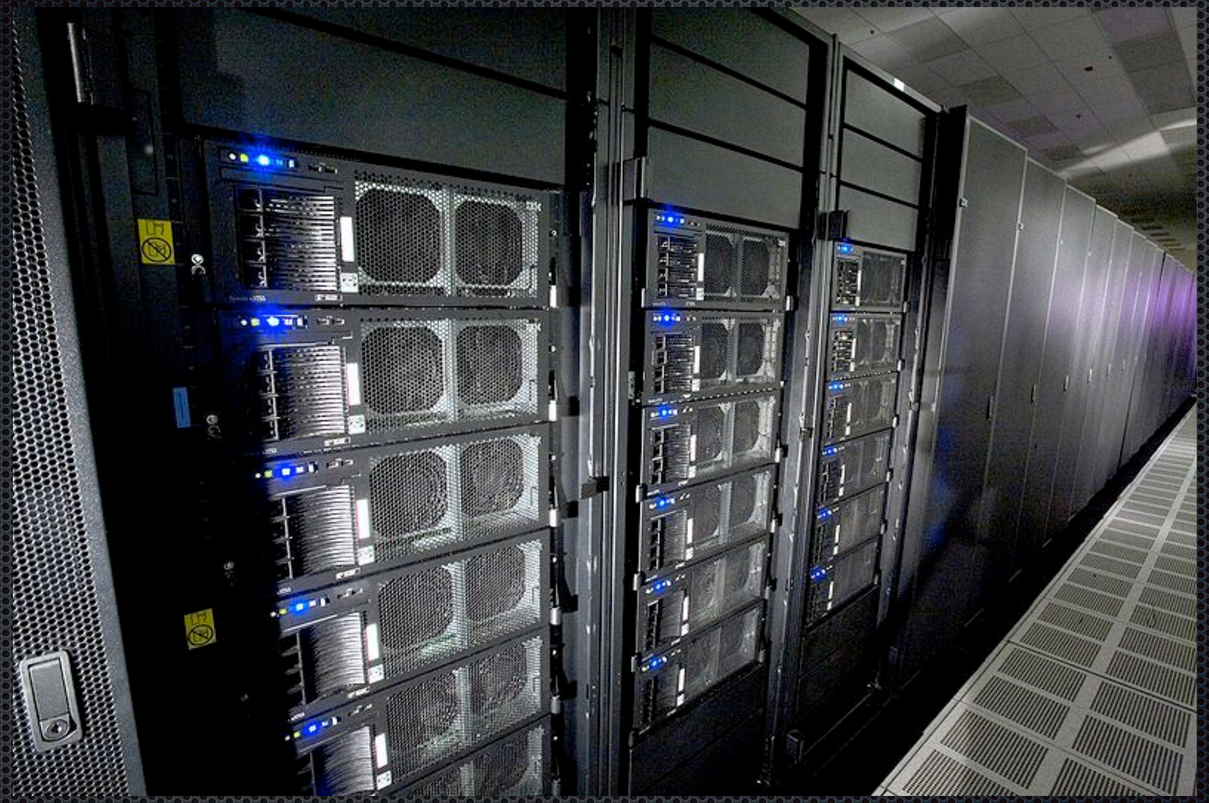
Red Storm

~13K quad-core Opterons, 284 TFLOPS,
Cray XT-4 (PPC 440-based network)

BlueGene L

596 TFLOPS, 75 TB memory, 1.9 PB disk
64K Dual PPC440, 3D Torus, Collective, Control

Roadrunner



~6500 dual Opteron + ~13K Cell + Infiniband
1st to reach 1 PFLOPS

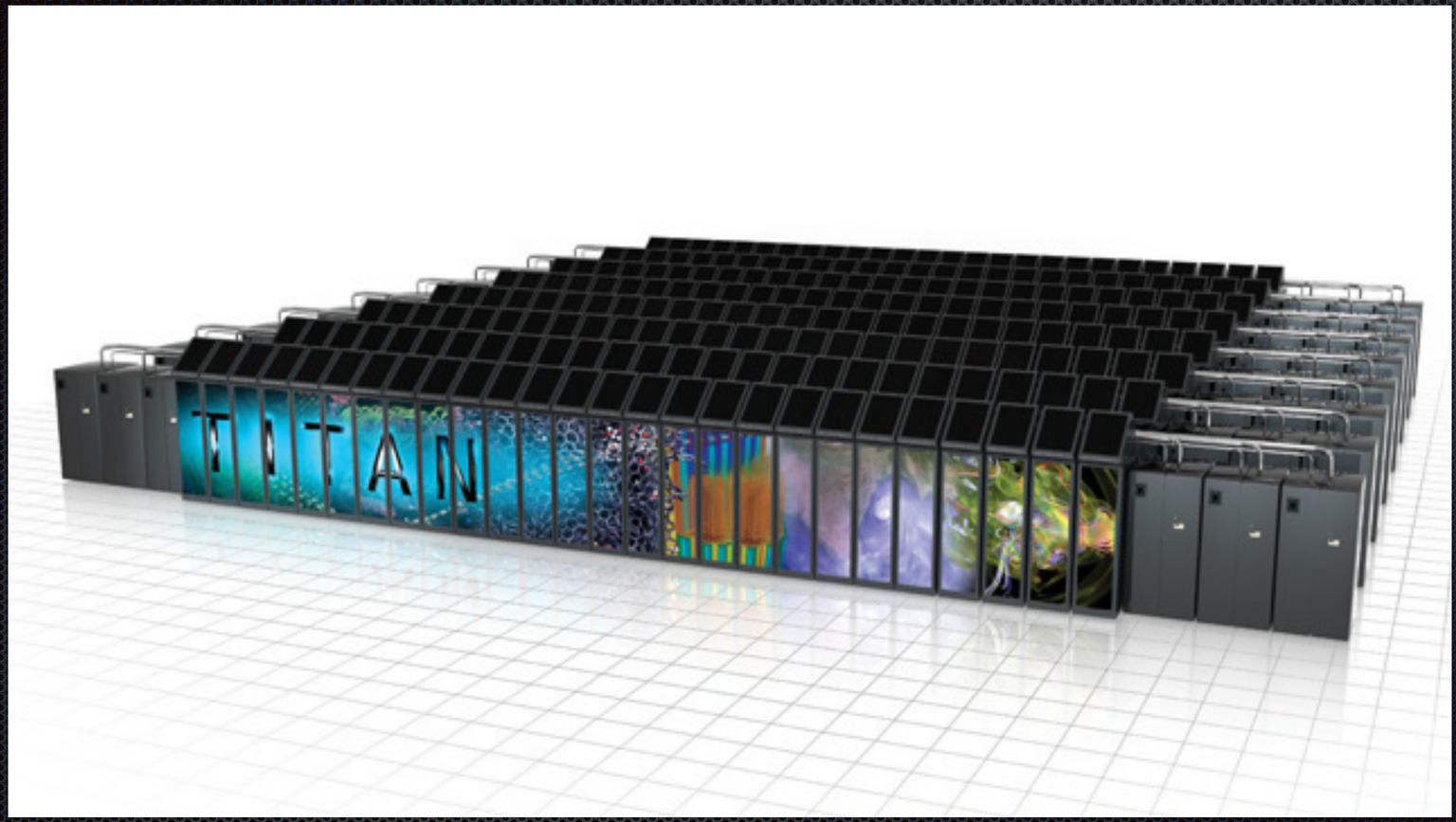
More Blue Gene

- Blue Gene/P: up to ~880K PPC450 @ 850 MHz, 3 PFLOPS
- Blue Gene/Q: (2011) 1.6M PPC cores, 6MW, 20 PFLOPS

Blue Gene Q

This one (ANL) 9PFLOPS, Sequoia (LLNL) 17PFLOPS
Each chip: 18 64-bit 4-way SMT PPC cores @1.6 GHz
96 racks, 98K nodes, 3K sq. ft. 7.9MW, 2GFLOPS/W

ORNL Titan



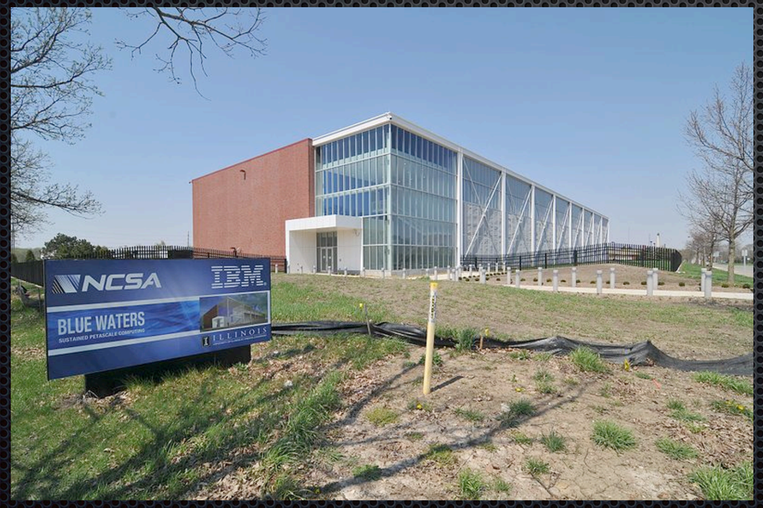
560,640 cores, plus GPU accelerators, 8.2 MW, 17.6 Petaflops

Tianhe-2

3.12 Million cores (Intel Ivy Bridge and Phi)
17.8 MW, 33.86 Petaflops

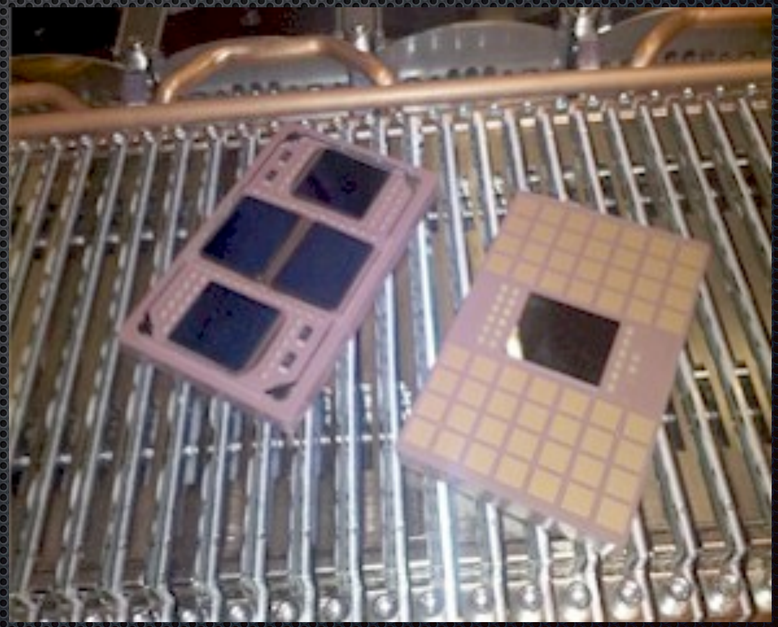
Blue Waters Facility

10 PFLOPS (peak) system being built at NCSA
(Univ. of Illinois)
90,000 sq ft building
20,000 sq ft machine room, w/6’ raised floor
32 MW power in 4 independent feeds



Blue Waters Processors

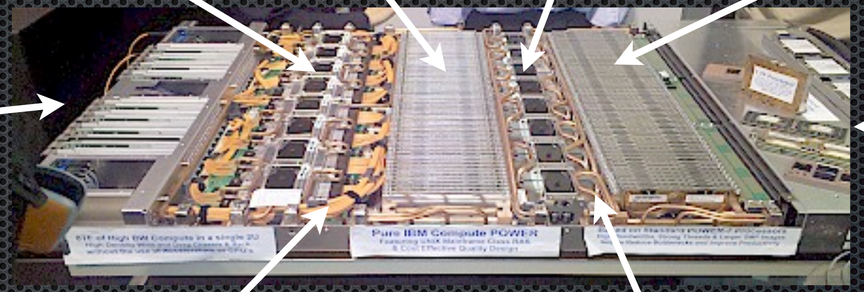
- Based on IBM Power 7
- 8-core, 32-thread, 4 GHz processor
- 32MB level 3 cache
- 2 Memory controllers, 8 channels
- 4 chips per module - 32 cores
- 1 TFlop per module
- 1.1 TB/s switch blocks



Blue Waters Drawers

- 8 processor modules (TFLOPS) per drawer
- 4 GB/core, 1 TB/drawer
- Water cooled
- 39” x 72” x 2”
- 290 lbs

Drawer Contents



Programming

- Initially MPI and Open-MP
- Migration to X10
  - Partitioned Global Address Space (PGAS)
  - Eclipse environment
  - Object-oriented, with ateach, foreach, atomic, clocks
  - Run on a virtual machine on top of HW

Muddied Waters

- Original design cancelled
  - IBM cannot deliver stated performance for bid price
  - NSF can’t pay more, IBM can’t dump hardware
- 235 Cray XE6 cabinets with AMD Opteron 6200
- 30 future Cray XK6 cabinets with NVIDIA Tesla GPUs

New Design

- 380,000 x86 cores + 3000 GPUs
- 3D torus interconnect (Cray Gemini - sold to Intel)
- 1.5 PB memory, 25 PB disk
- 11.5 PF peak performance

Exaflop Challenges

- 1 billion gigaflops, 1 million teraflops
- Traditionally 1 byte per flop - mostly memory machine
- No more clock scaling -- factor of 1000 beyond current petaflop systems all via additional parallelism
- Limited power and space (25 MW costs $15M/yr)
  - 25 picowatts per flop -- factor of 100 more efficient
- Heterogeneity, specialized processors, higher cost
- How do you checkpoint a thing like this?

Kevin Lim ISCA 2008

Understanding and Designing New Server Architectures for Emerging Warehouse-Computing Environments

Benchmark Suite

Table 1: Summary details of the new benchmark suite to represent internet sector workloads.

Workload	Emphasize	Description	Perf metric
websearch	the role of unstructured data	Open source Nutch-0.9, Tomcat 6 with clustering, and Apache2. 1.3GB index corresponding to 1.3 million indexed documents, 25% of index terms cached in memory. 2GB Java heap size. QoS requires >95% queries take <0.5 seconds.	Request-per-sec (RPS) w/ QoS
webmail	interactive internet services	Squirrelmail v1.4.9 with Apache2 and PHP4, Courier-IMAP v4.2 and Exim4.5. 1000 virtual users with 7GB of mail stored. Email/attachment sizes and usage patterns modeled after MS Exchange 2003 LoadSim heavy users. QoS requires >95% requests take <0.8 second.	RPS w/ QoS
ytube	the use of rich media	Modified SPECweb2005 Support workload with Youtube traffic characteristics. Apache2/Tomcat6 with Rock httpd server.	RPS w/ QoS
mapreduce	web as a platform	Hadoop v0.14 with 4 threads per CPU and 1.5GB Java heap size. We study two workloads - distributed file write (mapred-wr) and word count (mapred-wc)	Execution time

Cost/Performance in Server Farms

- Performance is measured in sustainable metrics, such as requests per second
- Total Cost of Ownership
  - Includes direct cost, power, cooling, depreciation
- Cost may be calculated in terms of TCO, power, etc.

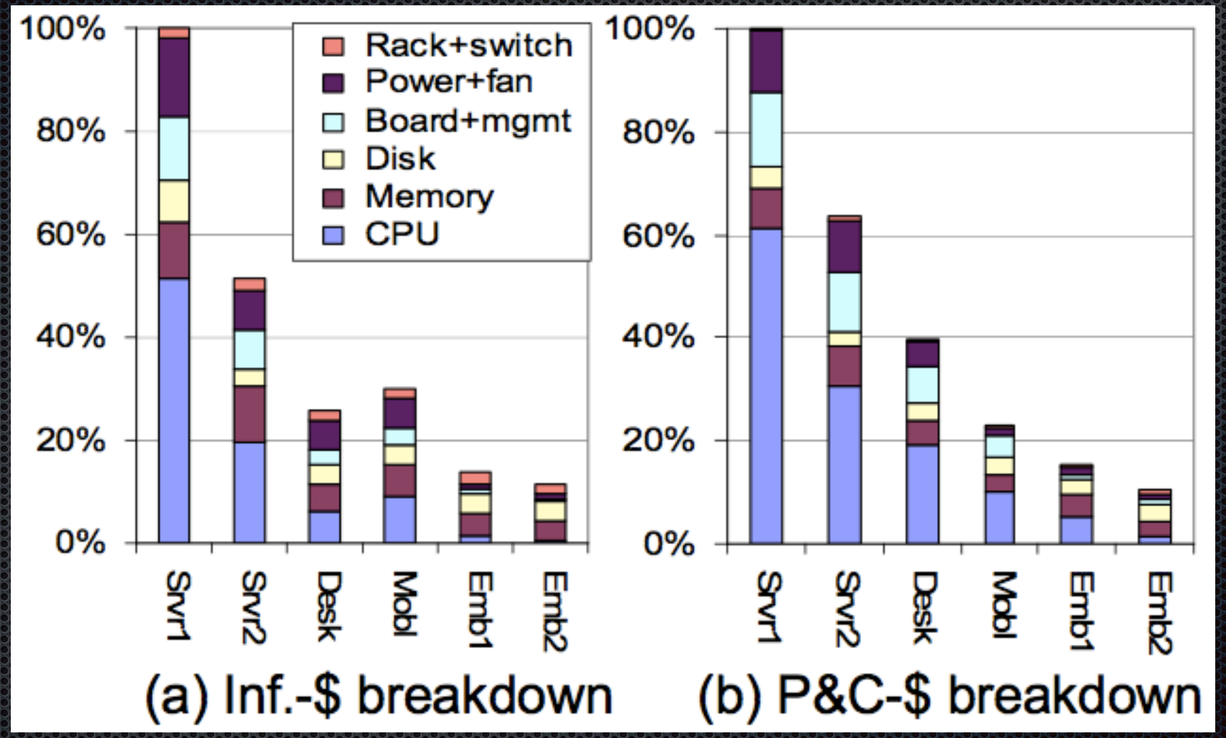
Systems Considered

- Infrastructure dollars

Table 2: Summary of systems considered.

System	"Similar to"	System Features	Watt	Inf-$
Srvr1	Xeon MP, Opteron MP	2p x 4 cores, 2.6 GHz, OoO, 64K/8MB L1/L2	340	3,294
Srvr2	Xeon, Opteron	1p x 4 cores, 2.6 GHz, OoO, 64K/8MB L1/L2	215	1,689
Desk	Core 2, Athlon 64	1p x 2 cores, 2.2 GHz, OoO, 32K/2MB L1/L2	135	849
Mobl	Core 2 Mobile, Turion	1p x 2 cores, 2.0 GHz, OoO, 32K/2MB L1/L2	78	989
Emb1	PA Semi, Emb. Athlon 64	1p x 2 cores, 1.2 GHz, OoO, 32K/1MB L1/L2	52	499
Emb2	AMD Geode, VIA Eden-N	1p x 1 cores, 600MHz, inord.,32K/128K L1/L2	35	379

Cost Breakdown



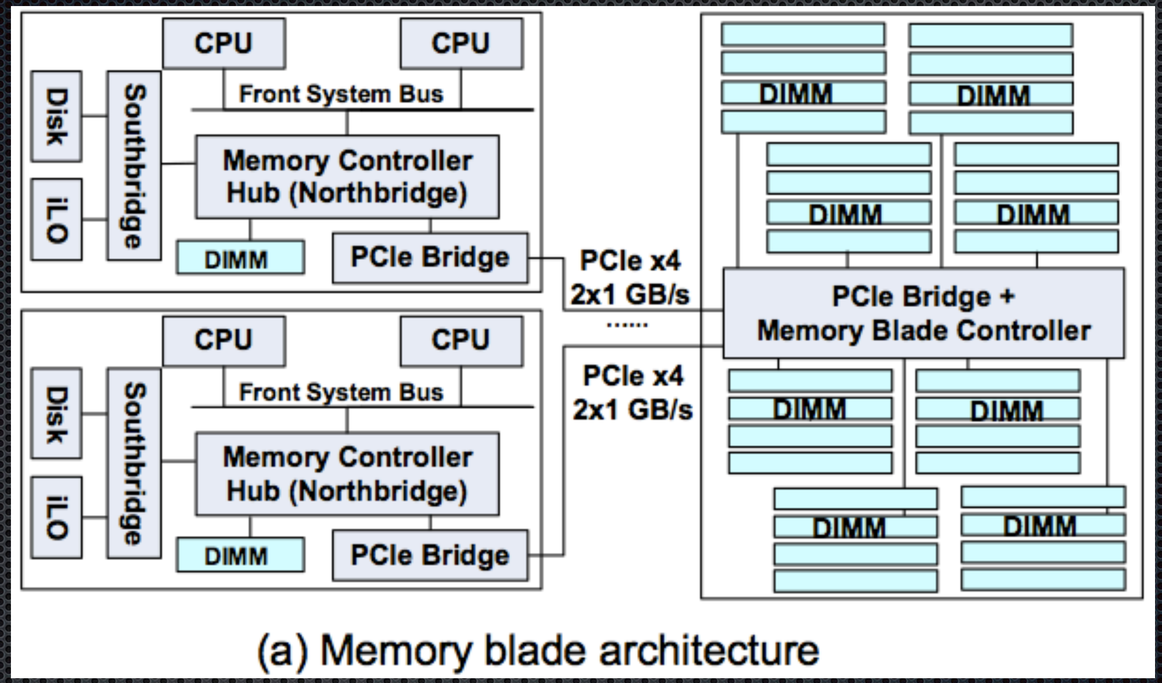
Infrastructure, Power and Cooling

Performance/Cost

	Workload	Srvr2	Desk	Mobl	Emb1	Emb2
Perf	websearch	68%	36%	34%	24%	11%
	webmail	48%	19%	17%	11%	5%
	ytube	97%	92%	95%	86%	24%
	mapred-wc	93%	78%	72%	51%	12%
	mapred-wr	72%	70%	54%	48%	16%
	HMean	71%	42%	38%	27%	10%
Perf/Inf-$	websearch	133%	139%	112%	175%	93%
	webmail	95%	72%	55%	83%	44%
	ytube	188%	358%	315%	629%	206%
	mapred-wc	181%	302%	241%	376%	101%
	mapred-wr	141%	272%	179%	350%	140%
	Hmean	139%	162%	125%	201%	91%
Perf/W	websearch	107%	90%	147%	157%	103%
	webmail	76%	47%	73%	75%	49%
	ytube	152%	233%	413%	566%	229%
	mapred-wc	146%	197%	315%	338%	113%
	mapred-wr	114%	177%	235%	315%	157%
	Hmean	112%	105%	164%	181%	101%
Perf/TCO-$	websearch	120%	113%	124%	167%	97%
	webmail	86%	59%	62%	80%	46%
	ytube	171%	291%	351%	600%	215%
	mapred-wc	164%	246%	268%	359%	106%
	mapred-wr	128%	221%	200%	334%	147%
	Hmean	126%	132%	140%	192%	95%

(c) Performance, cost and power efficiencies

Proposed 2-Level Memory Architecture

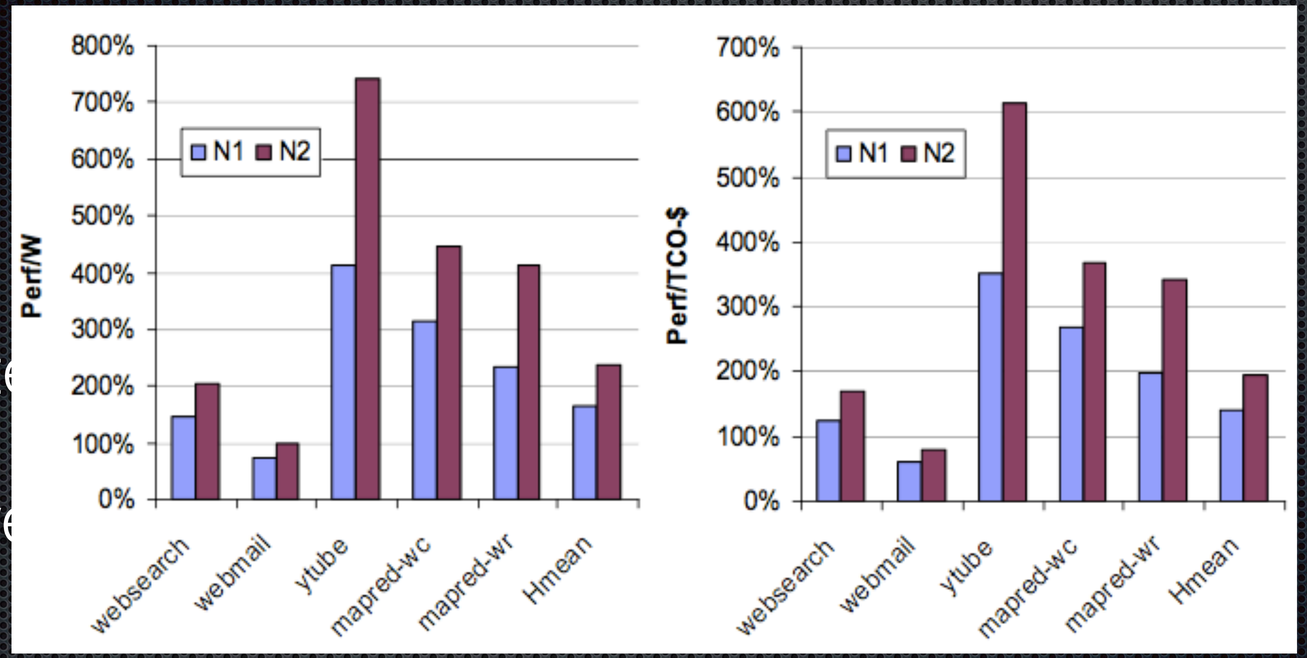


(a) Memory blade architecture

Reduce amount of DRAM by sharing
est. 8% to 11% better perf/$TCO

Integrated Systems

- N1 is current te
- N2 adds 2-leve



Discussion

Svilen Kanev ISCA 15

Profiling a warehouse-scale computer

Some Observations

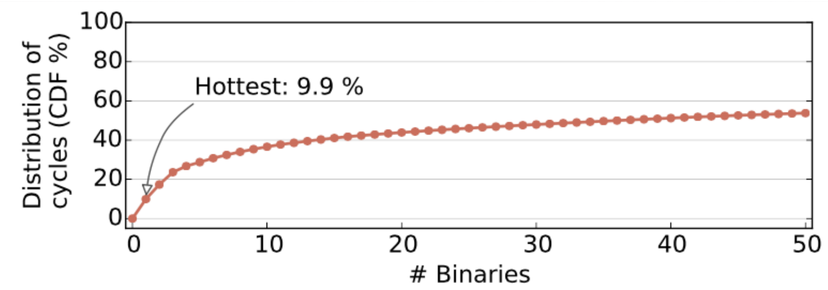


Figure 1: There is no "killer application" to optimize for. The top 50 hottest binaries only cover ≈60% of WSC cycles.

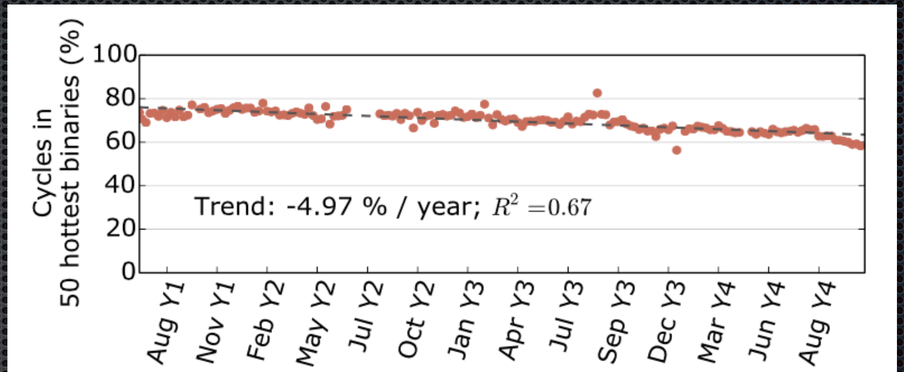


Figure 2: Workloads are getting more diverse. Fraction of cycles spent in top 50 hottest binaries is decreasing.

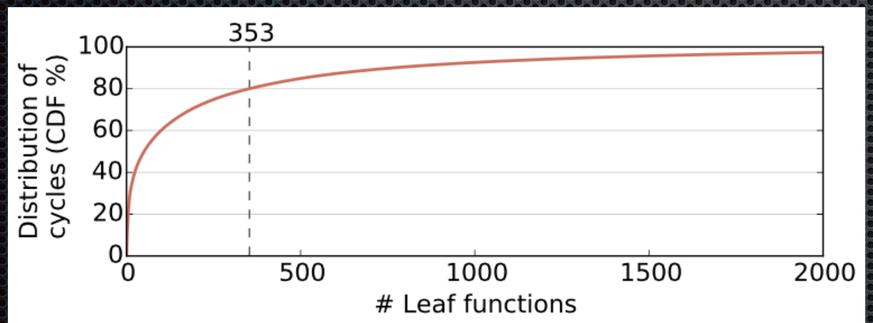


Figure 3: Individual binaries are already optimized. Example binary without hotspots, and with a very flat execution profile.

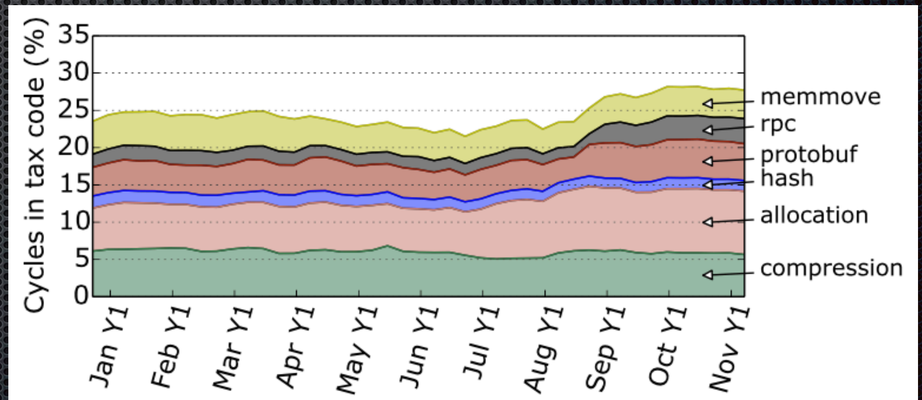


Figure 4: 22-27% of WSC cycles are spent in different components of "datacenter tax".

More Observations

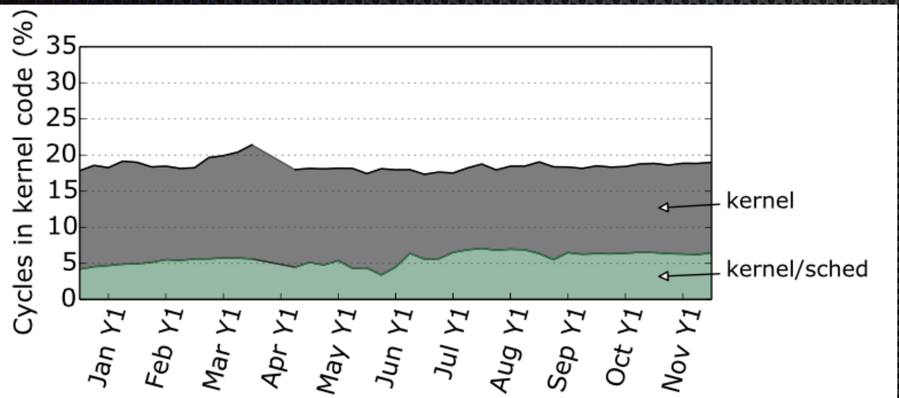


Figure 5: Kernel time, especially time spent in the scheduler, is a significant fraction of WSC cycles.

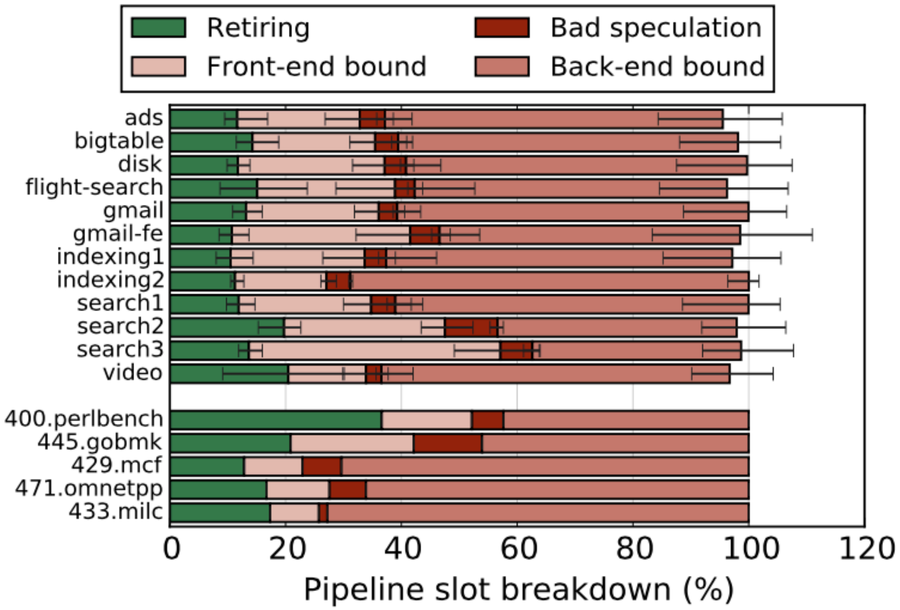


Figure 6: Top-level bottleneck breakdown. SPEC CPU2006 benchmarks do not exhibit the combination of low retirement rates and high front-end boundedness of WSC ones.

Benchmarks like SPEC are not representative

Caches

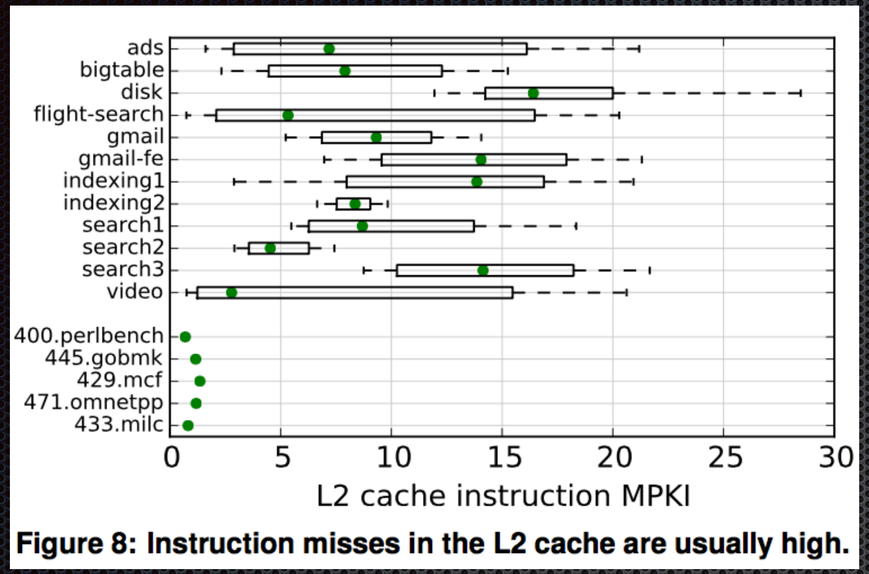


Figure 8: Instruction misses in the L2 cache are usually high.

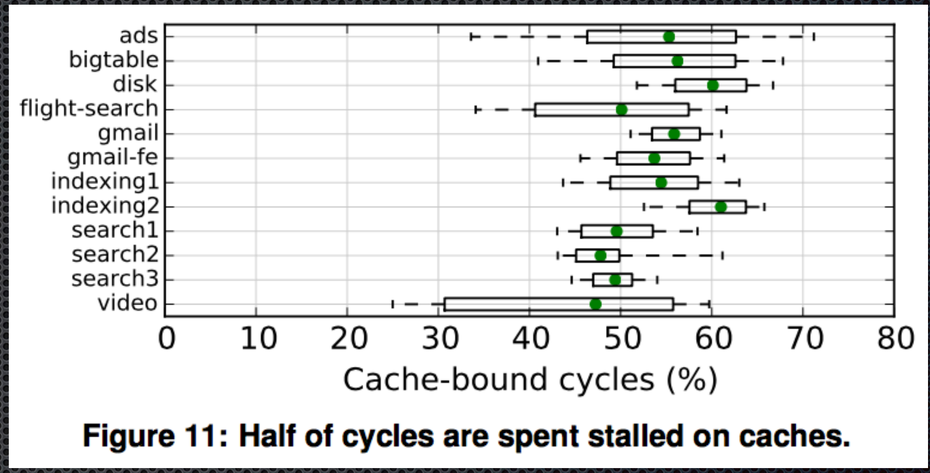


Figure 11: Half of cycles are spent stalled on caches.

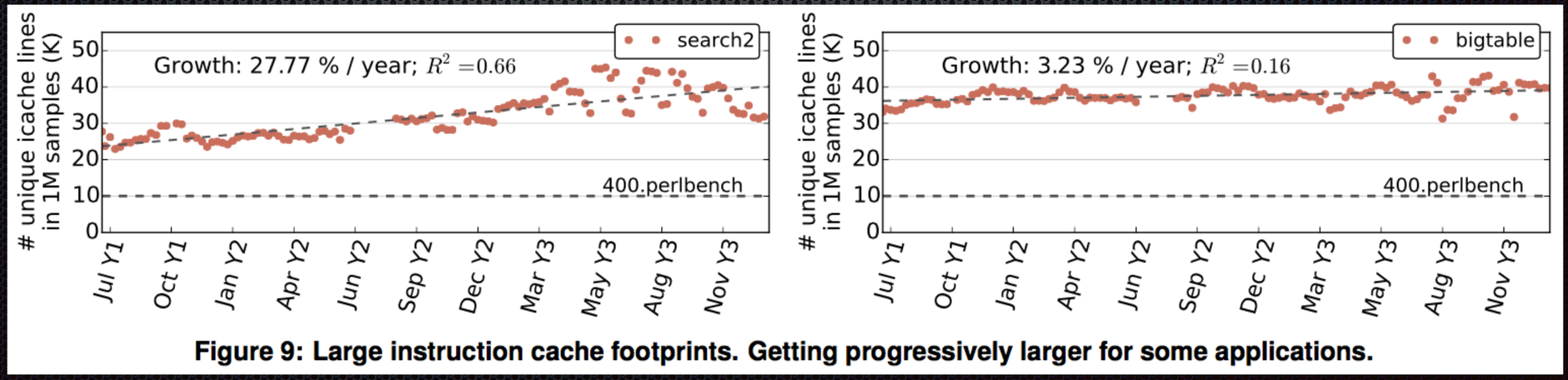


Figure 9: Large instruction cache footprints. Getting progressively larger for some applications.

SuperScalar

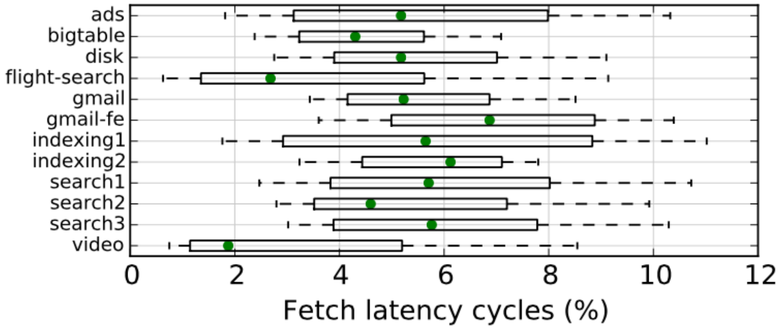
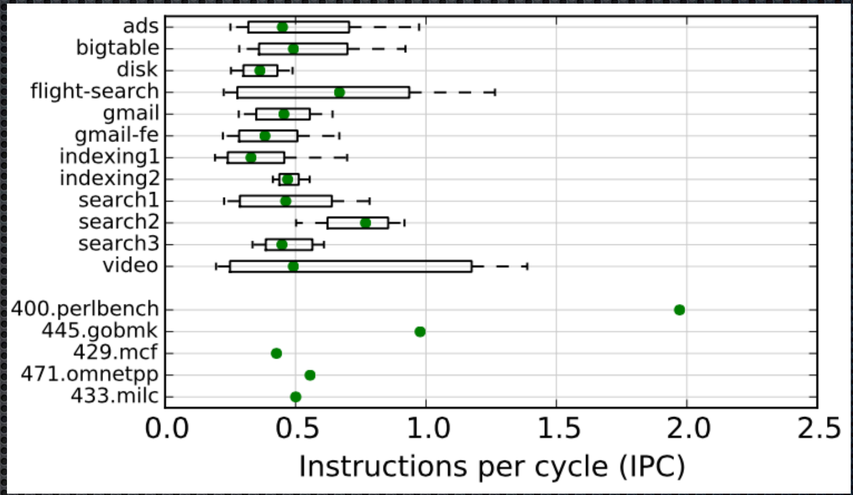


Figure 7: Cycles completely starved on front-end bottlenecks account for more than 5% of execution.



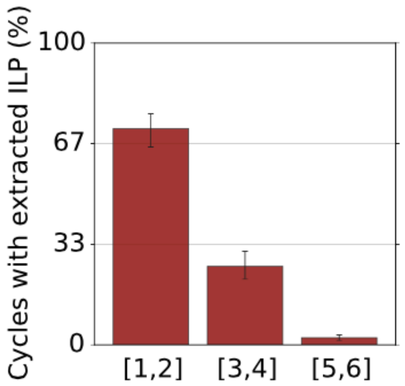


Figure 12: Extracted ILP. 28% of cycles utilize 3 or more execution ports on a 6-wide machine.

Multithreading

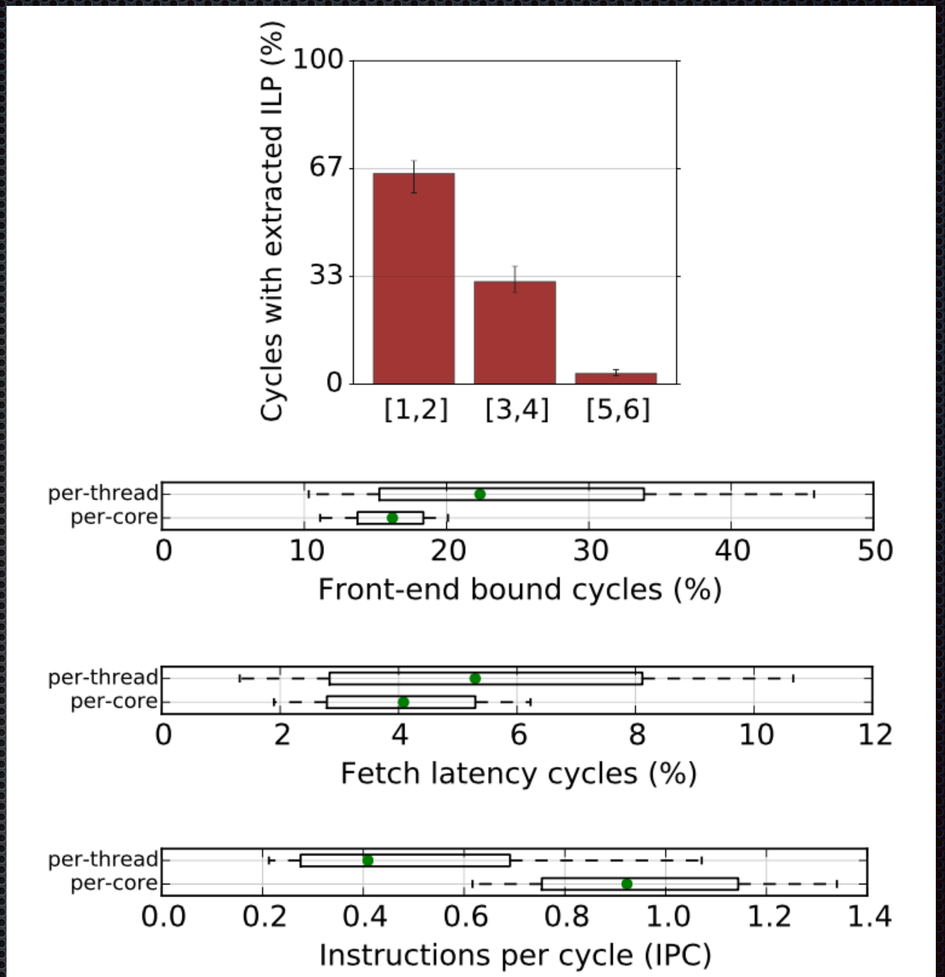


Figure 14: SMT effects on architectural behavior. From top to bottom: (i) more ILP extracted compared to Figure 12; (ii) front-end bound cycles decrease, but (iii) instruction starvation still exists; (iv) core throughput doubles with two hyperthreads.

Conclusions

Finding	Investigation direction
`workload diversity`	Profiling across applications.
`flat profiles`	Optimize low-level system functions.
`datacenter tax`	Datacenter specific SoCs (protobuf, RPC, compression HW).
`large (growing) i-cache footprints`	I-prefetchers, i/d-cache partitioning.
`bimodal ILP`	Not too “wimpy” cores.
`low bandwidth utilization`	Trade off memory bandwidth for cores. Do not use SPECrate.
`latency-bound performance`	Wider SMT.

Summary of findings and suggestions for future investigation.

Discussion