Ronald Luijten – Data Motion Architect
lui@zurich.ibm.com

IBM Research - Zurich

14 March 2012

Co-design perspectives from a Data Motion Architect

DISCLAIMER

This presentation is entirely Ronald's view and not necessarily that of IBM.

acknowledgement

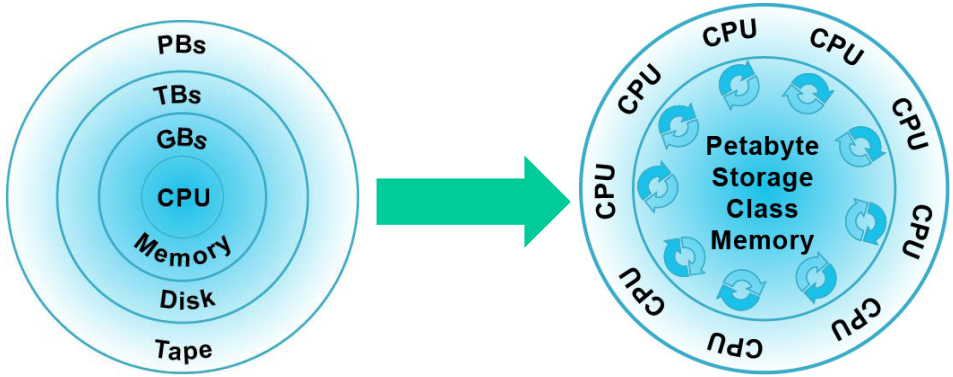
- I work with a great team
- Many of the charts in this preso are from **Phillip Stanley-Marbell** and **Victoria Caparros**

About you

SOS in a BOX

DCO team 3 year horizon

- Datacenter disruption ahead due to 5 Walls, SCM and 3D packaging
- The datacenter challenge is shifting from compute to data
  - MemcacheD; Purescale; Text Analytics
  - PByte scale memory servers needed
  - Single address space, not cache coherent
  - Small access granularity (few bytes)
  - Cache unfriendly access patterns
    - Large scatter / gather



- Need I/O architecture that satisfies above
  - Address translation bottleneck
  - Power (P9) I/O architecture mission

- Need new, holistic modeling methodology to allow design space exploration

Perspectives on Co-design

- Not co-optimization but holistic optimization is needed

- It was mentioned a few times yesterday:
  - need to holistically select physics model, numerical methods, algorithm, programming model and hardware

- Moore's paradox was mentioned yesterday:
  - Cost of DRAM bit, BW becomes larger wrt. Cost of compute
- I mentioned this the last two SOS meetings: Data is the problem – compute has been solved

- We are developing a design space exploration environment, where aim to capture from algorithm to transistor based on first principles
  - High risk – high potential payoff endeavor

- We are establishing new insights between SW and HW (examples follow)

- For exascale, all components / building blocks are understood today – however, not how they are to be put together!
  - This will be dictated by energy constraints

“core is the most important part of the system”

- Do whatever you can to keep it busy
  - Utilization is paramount, don’t care how much area and energy is spent doing that.
  - Speculation is good – in case it works, you gain, in case it does not work, nothing is lost, right?

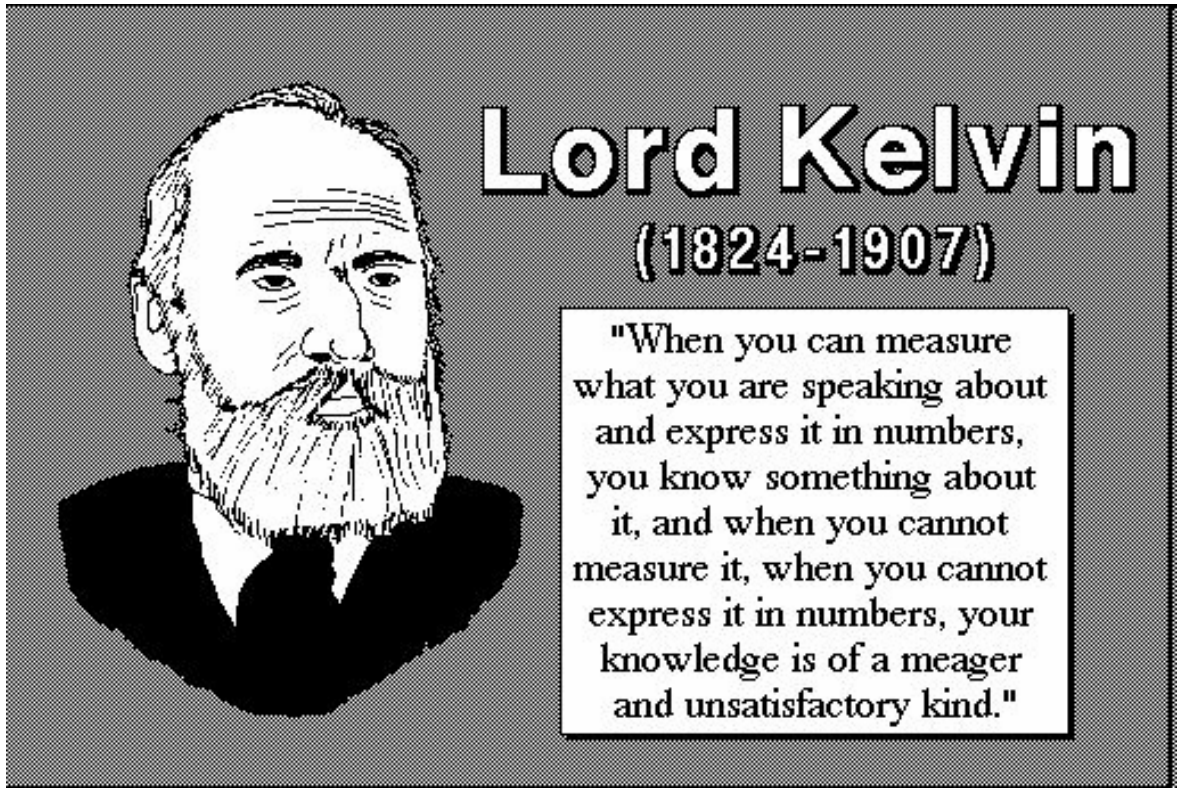
Moving Data is a sin

- Avoid data motion as much as possible
- Compute is free – put it where ever it needs to be done – even when utilization is almost zero

On Data Locality:

- Temporal
- Spatial
- Geographical

You get what you measure (Lord Kelvin)



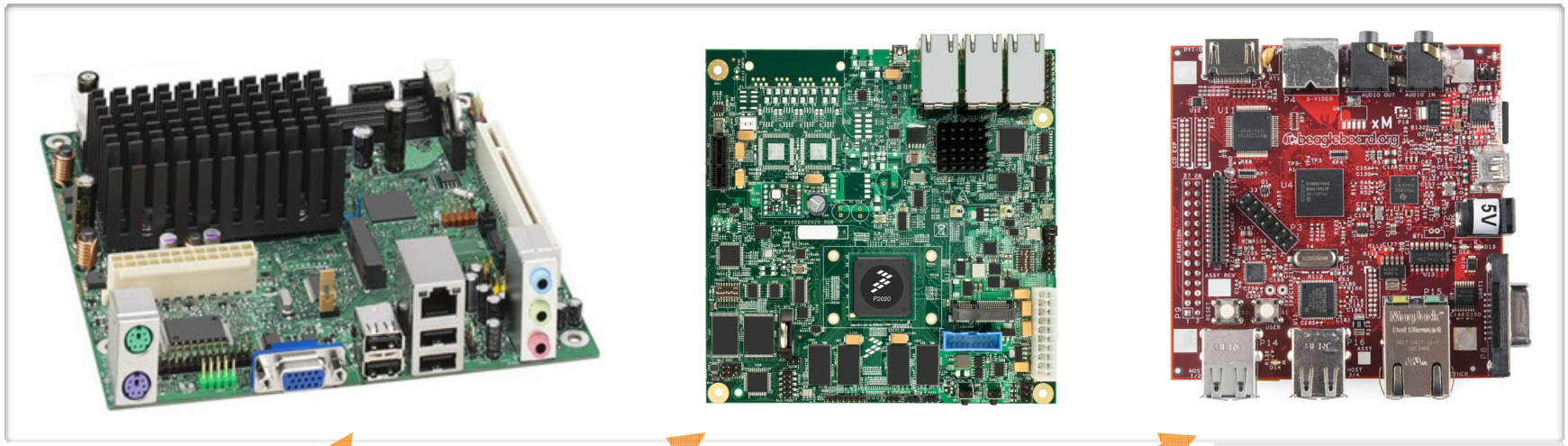
- Corollary:
  Don't expect to get what you don't measure

- Key is to decide what to measure (seems trivial, right?)

Ronald's "More's law" – the result of Moore's law

Performance and Energy measurements in 3 ISAs

- Candidate processors for scale-out systems: Atom, PowerPC, ARM
  - Three hardware reference designs; all processors implemented in 45nm CMOS
  - All three systems running Linux distributions based on kernel 2.6.32

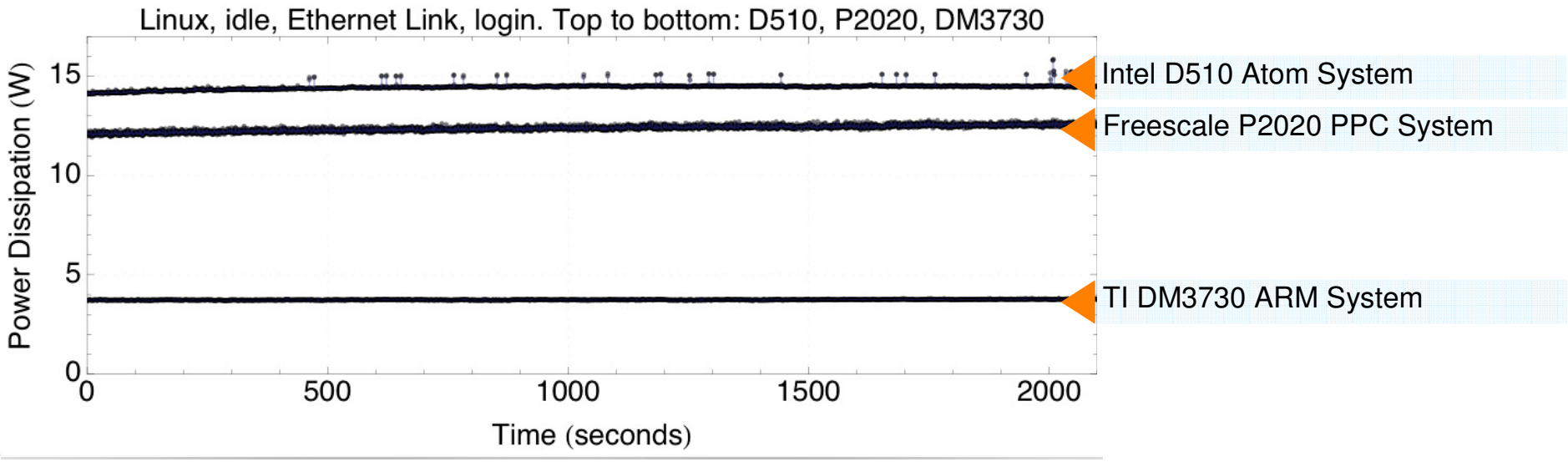


Platform	General-Purpose Cores	CPU Clock	Cache Hierarchy
Atom™ D510MO	2 x86-64	1.66 GHz	32 K/24 K L1 I/D (per core) and 1 M L2
Freescale™ P2020RDB	2 Power Architecture® e500	1.0 GHz	32 K/32 K L1 I/D (per core), 512 K shared L2
TI DM3730 (Beagleboard-xM)	1 ARM® Cortex™-A8	1.0 GHz	32 K/32 K L1 I/D, 256 K L2

Pictures are NOT to scale

Common Properties:
- Identical 4GB flash disk
- Lab-grade power meter, 1 measurement per second, sub-mA resolution
- Power measured for whole platform

Zero-Load System Whole-System Power Analysis



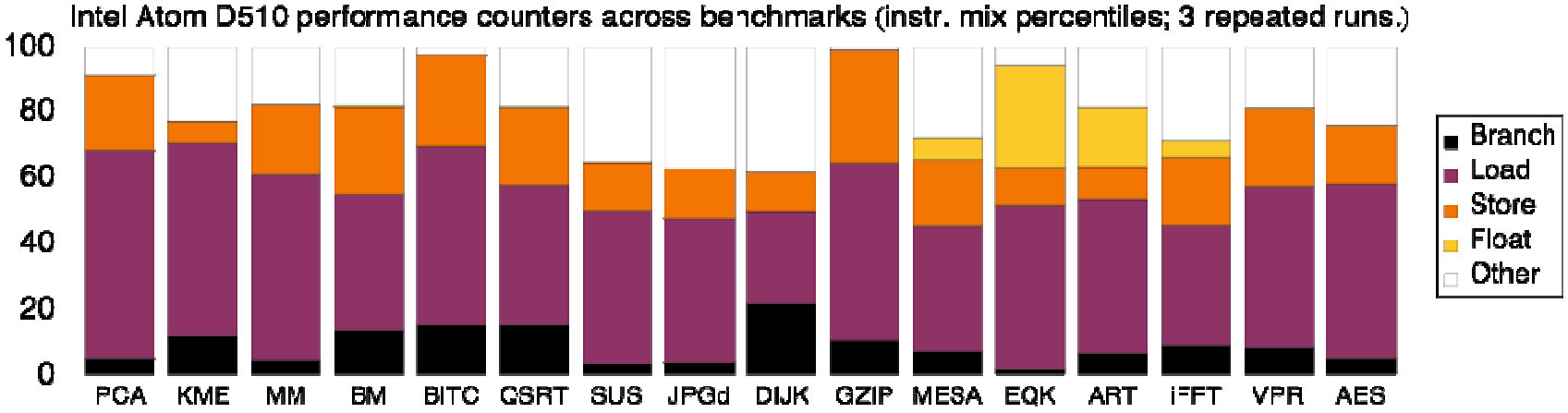
- Linux running, no apps – idle process
- Significant difference in zero-load power dissipation
  - > 3× difference in zero-load power dissipation, from lowest (3.8W), to highest (14.5W)

Workload

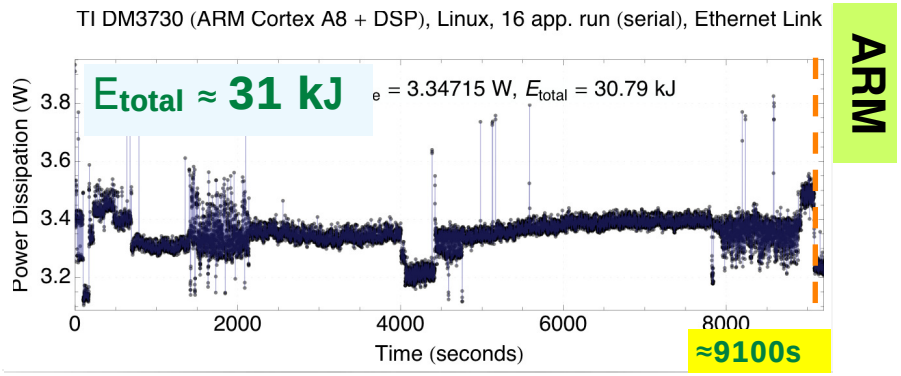
- 16 applications
  - Broad range of domains
  - Range from small data-analytics kernels with poor locality (**KME**), to large compute-intensive applications (**ART**, **EQK**)

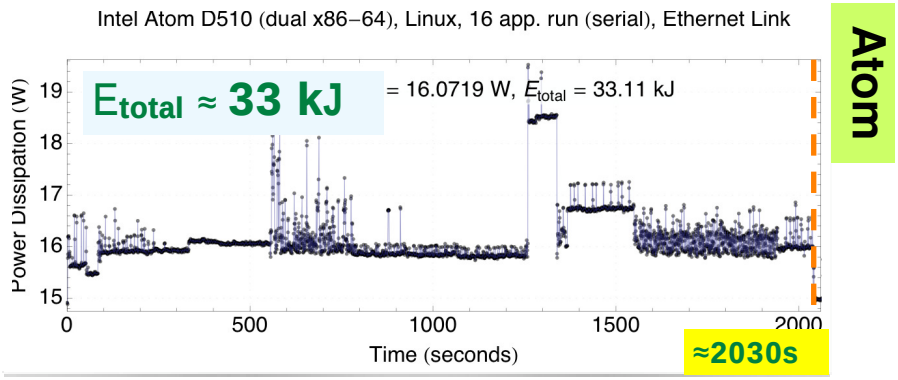
Application	Benchmark Suite	Dwarf
kmeans (KME)	**Phoenix MapReduce**	Dense Linear Algebra
matrix multiply (MM)	—	Sparse Linear Algebra
pca (PCA)	—	Dense Linear Algebra
basicmath (BM)	**MiBench**	Sparse Linear Algebra
bitcount (BIT)	—	Sparse Linear Algebra
dijkstra (DIJK)	—	Graph Traversal
JPEG decode (JPGd)	—	Dense Linear Algebra, Structured Grids
iFFT (iFFT)	—	Spectral Methods
qsort (QSRT)	—	Graph Traversal
rijndael decode (AES)	—	Combinational Logic
susan (SUS)	—	Dense Linear Algebra
164.gzip (GZIP)	**SPEC CPU2000**	Finite State Machine
175.vpr (VPR)	—	Backtrack/Branch+Bound
177.mesa (MESA)	—	MapReduce
179.art (ART)	—	Backtrack/Branch+Bound
183.equake (EQK)	—	Unstructured Grids

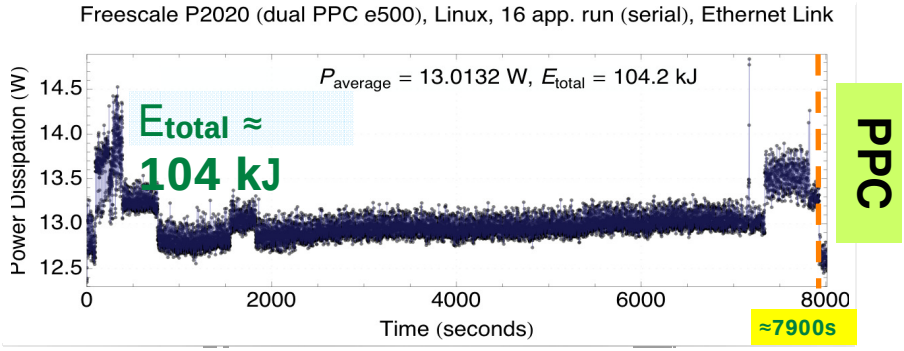
- Instruction mixes:



Application-Driven Whole-System Power Analysis

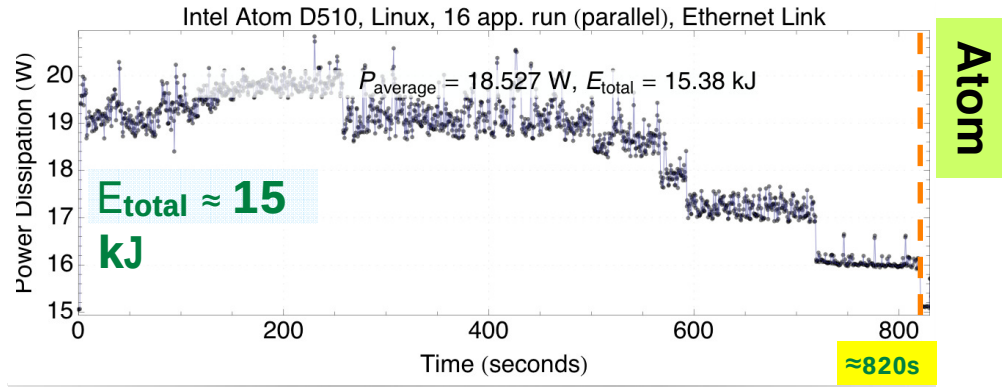


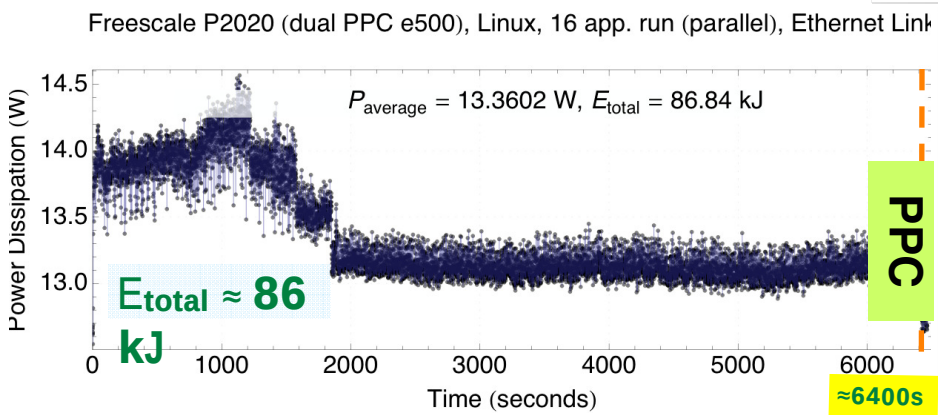




- Serial workload
  - Single core, serial application launch
  - Essentially tied between ARM (30.8kJ) and Atom (33.1kJ): 7.8% difference
  - PPC: due to FPU

Application-Driven Whole-System Power Analysis





- "Throughput" workload
  - All 16 applications launched simultaneously; on Atom and PPC, utilize 2 cores
  - PPC: has limited FPU; perf. limited by floating-point-intensive **Equake** benchmark
- Highest-average-power system is the most energy-efficient
  - Atom platform uses least energy (15.4kJ), vs. ARM (30.8kJ) and PPC (86.8kJ)

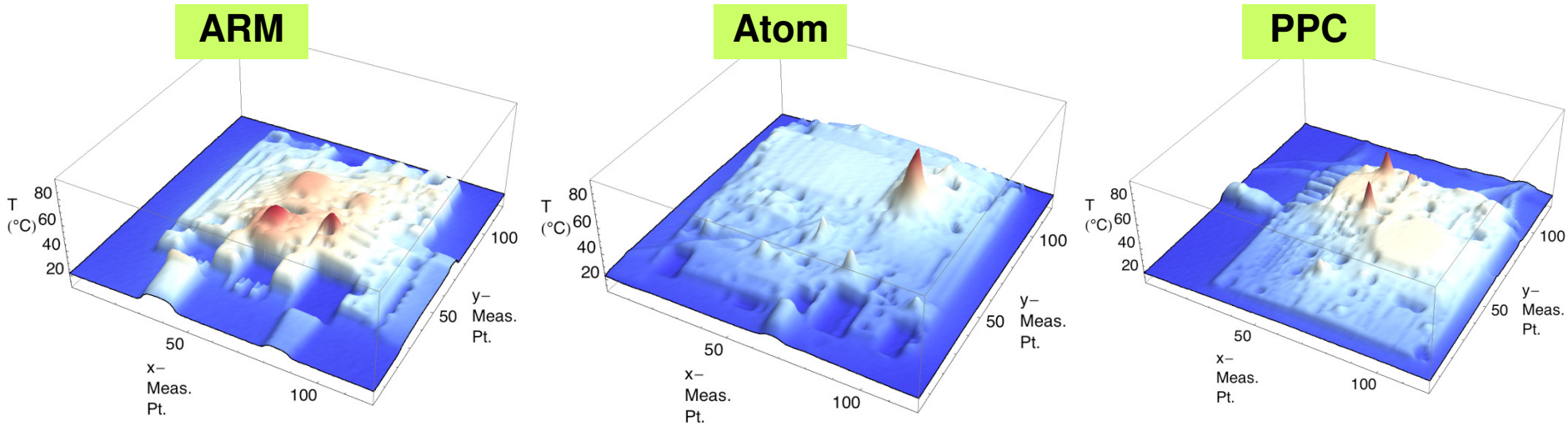
Thermal Analysis and Board-Level Power Breakdowns

Stefan-Boltzmann Law: relates energy flux (Φ) to temperature (T), area (A), and emissivity (e), Stefan-Boltzmann constant (σ **= 5.67 × 10^{-8}Wm^{-2}K^{-4}**):

$$\Phi = A \cdot e \cdot \sigma \cdot T^4$$

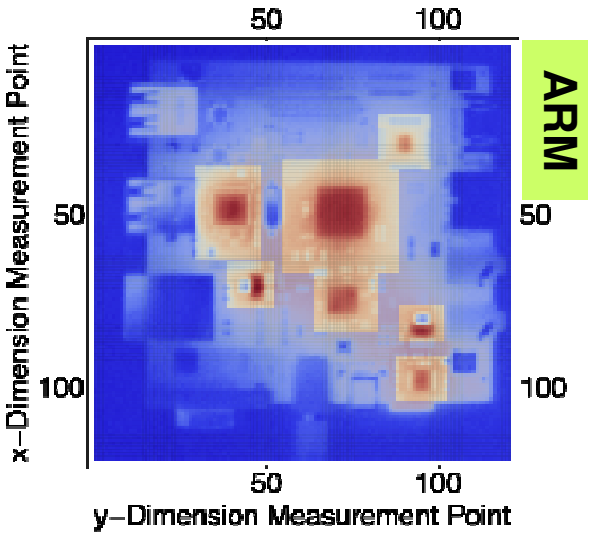
Thermal Analysis: Hotspots

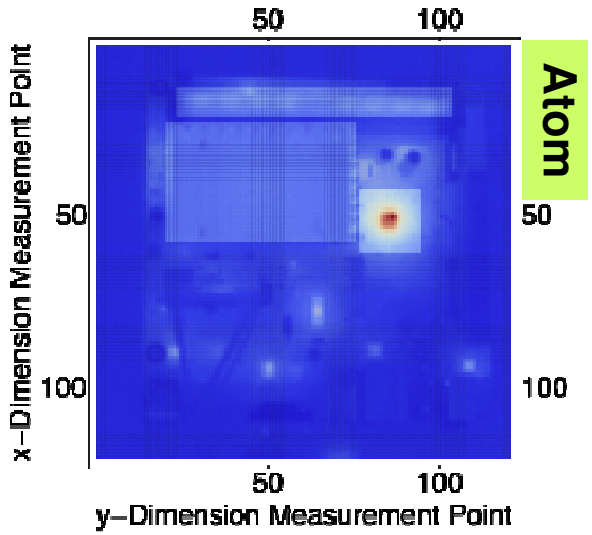
2D temperature map from **14k-element microbolometer array** (thermal radiation meter)

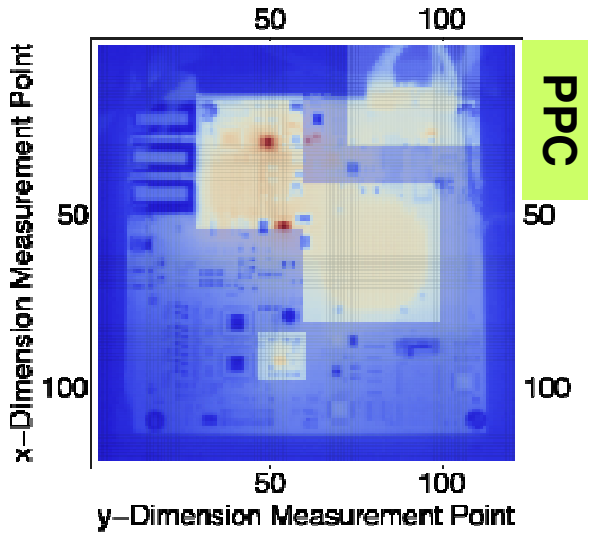


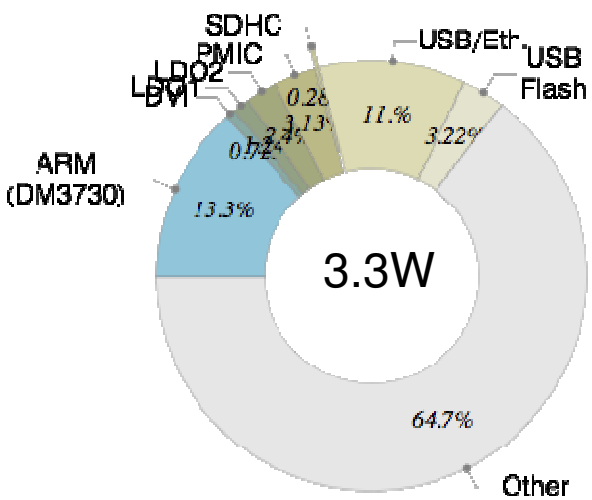
- Temperature peaks
  - Hottest components in all three platforms are not the processors
  - **ARM platform:** USB-Ethernet bridge;
  - **Atom platform:** I/O controller;
  - **PowerPC platform:** Ethernet switch subsystem

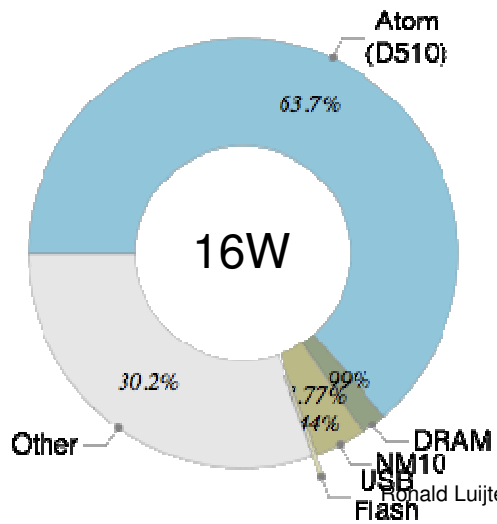
Thermal Analysis: Estimating Power Apportionment

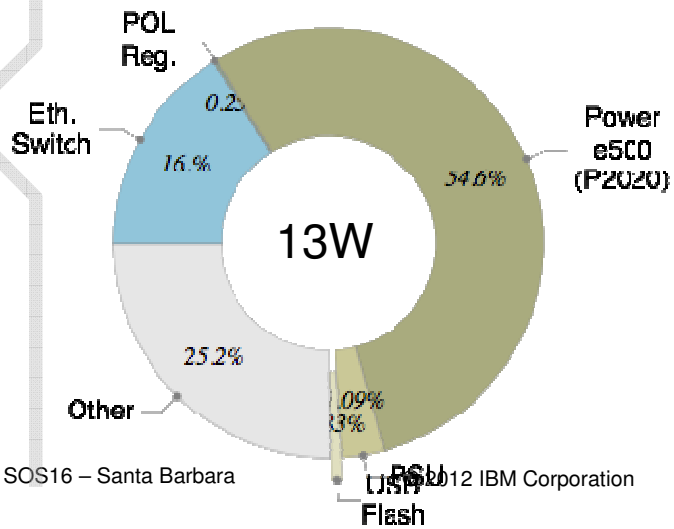




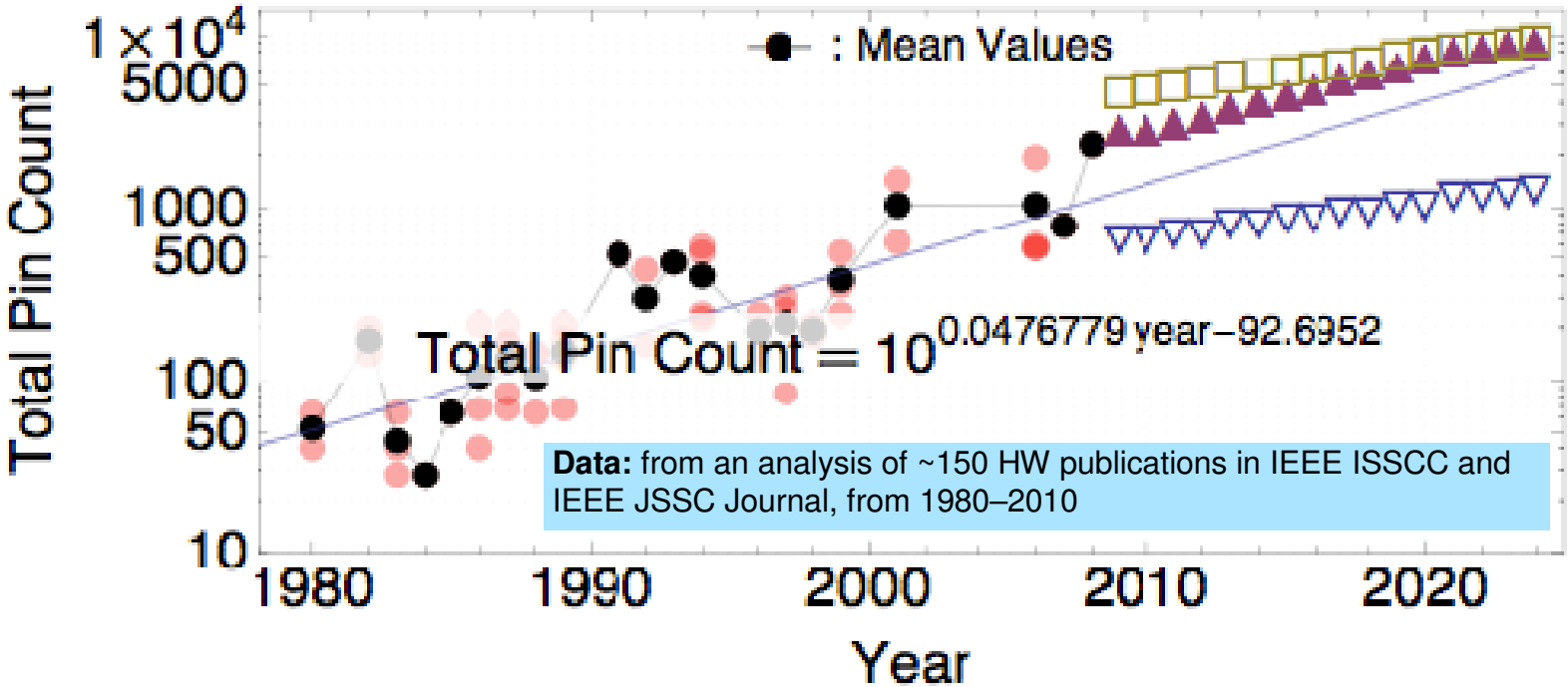








Power wall analysis

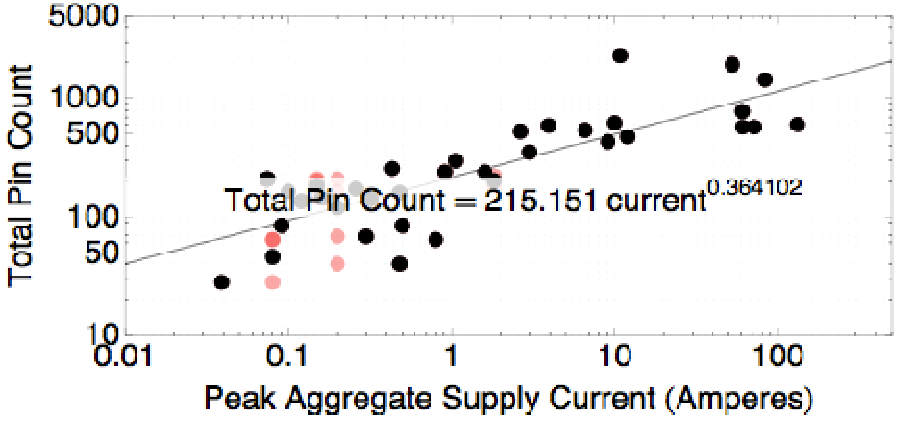


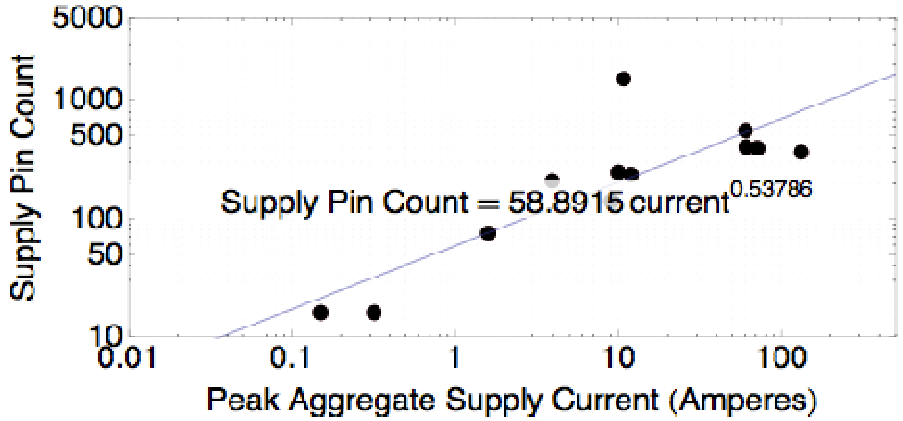
ITRS projections: ▼: low-end; ▲ high-end: ■: high-performance microprocessors and ASICs

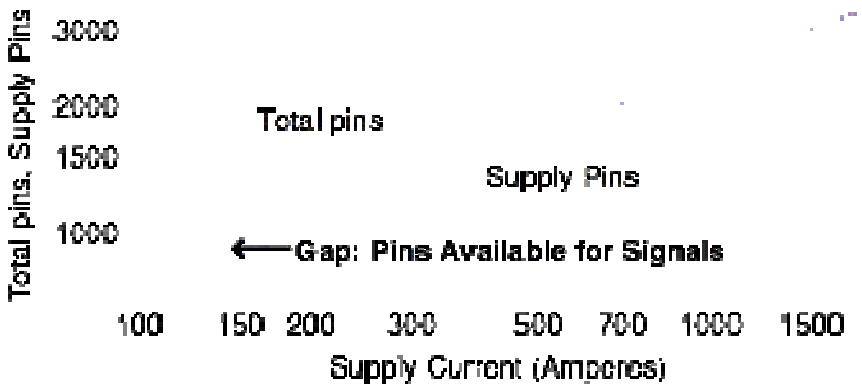
Assumptions:

- Keep scaling as we have done in past couple of years
- Projection from 30 year trend until 2010
- ITRS roadmap on pins

Power wall analysis: Supply pins





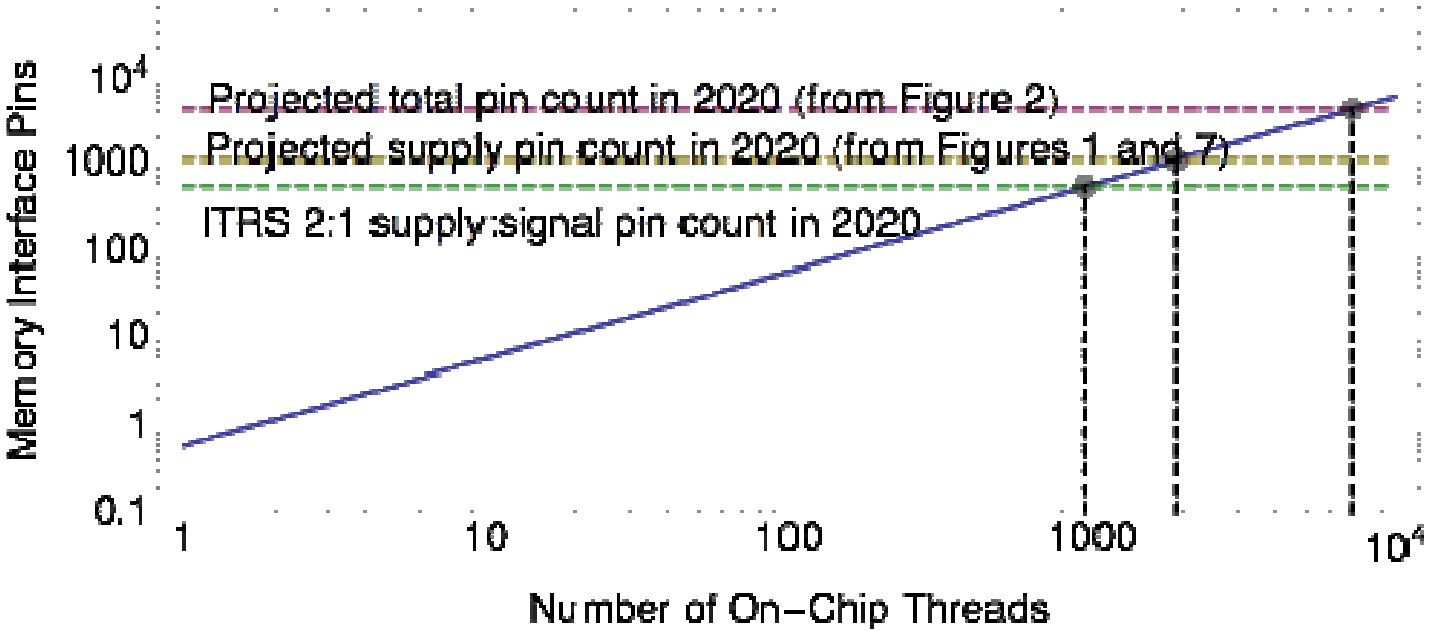


Projection based on historical data

- Contradicts ITRS assumption

1:1 for ASIC, 2:1 for CPUs – supply:signal)

Power—Interplay Between Power and Bandwidth: Pins



Assumptions:

- 17GB/s mem @ 155 pins (DDR3)
- 1 memory acc / 1k instr / 32 byte fetch
- ITRS pin projections for 2020

If we keep scaling as we have in past, we can support ~1000 threads per chip in 2020

P. Stanley-Marbell, V. Caparrós Cabezas, and R. Luijten **"Pinned to the Walls—Impact of Packaging on the Memory and Power Walls"**, IEEE/ACM International Symposium on Low-Power Electronics and Design (ISLPED 2011).

Conclusions

- Energy: ISA does not matter – system design does
  - Packaging is *the* make or break dimension for exascale
- No ARM ISA energy advantage measured – contradictory to rumors
  - The u-server advantage lies in:
    - Simple core leaves room to integrate 'other stuff' onto same chip
    - Ethernet, USB, SATA, etc. integrated onto chip
    - Significant energy savings by avoiding chip-crossings (System-on-a-Chip)
- Need to understand workloads, data placement, access patterns in order to optimize system design
- Choice of metrics is key – exaflop as target is missing the point
- Need to understand energy efficient cache design
- Creating holistic design space exploration tool
- Breakthru Innovation needed in system design
  - Start from data – work your way back to processing
  - Putting 100's of cores on single die most likely wrong design point for HPC
  - I predict accelerators will be key – not sure this will be current GPU thinking

papers (µServer research)

- **“Parallelism and Data Movement Characterization of contemporary Application Classes** ”, Victoria Caparros Cabezas, Phillip Stanley-Marbell, to appear in ACM SPAA 2011, June 201
- “**Quantitative Analysis of the Berkeley Dwarfs' Parallelism and Data Movement Properties**”, Victoria Caparros Cabezas, Phillip Stanley-marbell, to appear in ACM CF 2011, May 2011
- “**Performance, Power, and Thermal Analysis of Low-Power Processors for Scale-Out Systems**”, Phillip Stanley-Marbell, Victoria Caparros Cabezas, IEEE HPPAC 2011, May 2011
- “**Pinned to the Walls—Impact of Packaging and Application Properties on the Memory and Power Walls**”, Phillip Stanley-Marbell, Victoria Caparros Cabezas, Ronald P. Luijten, IEEE ISLPED 2011, Aug 2011.