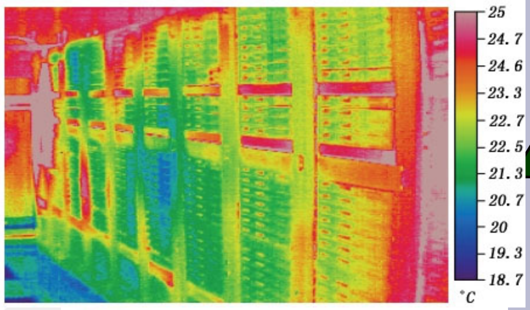
Power-Metering in Virtualized Datacenters

Ada Gavrilovska, Bhavani Krishnan, Hrishikesh Amur,
Karsten Schwan, Surabhi Diwan,
Matthew Wolf, Jhenkar Vidyashankar, Hui Chen, …
Hsien-Hsin Lee, Eric Fontaine

Green Computing Initiative

http://img.all2all.net/main.php?g2_itemId=157

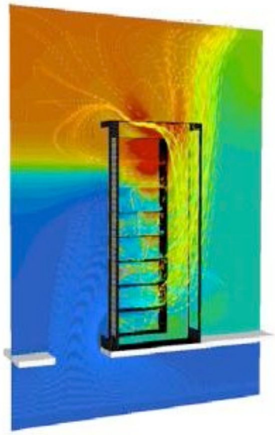


focus of our work:

Datacenter and beyond: design, IT management, HVAC control… (ME, SCS, OIT…)

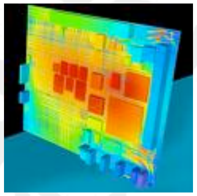
↕

Rack: mechanical design, thermal and airflow analysis, VPTokens, OS and management (ME, SCS)



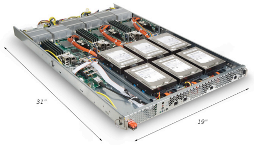
↕

Board: VirtualPower, scheduling/ scaling/operating system… (SCS, ME, ECE)



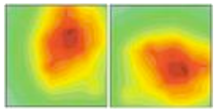
↕

Chip and Package: power multiplexing, spatiotemporal migration (SCS, ECE)



↕

Circuit level: DVFS, power states, clock gating (ECE)



Power distribution and delivery (ECE)

Power-aware Datacenter Management

- Continuous power monitoring
  - RPDUs
  - SNMP or IPMI based infrastructure
- Continuous resource usage monitoring
  - Ganglia, SNMP, or EVPath based
  - aggregate and per VM usage of CPU, Mm, IO…
- Dynamic load reconfiguration
  - *???*

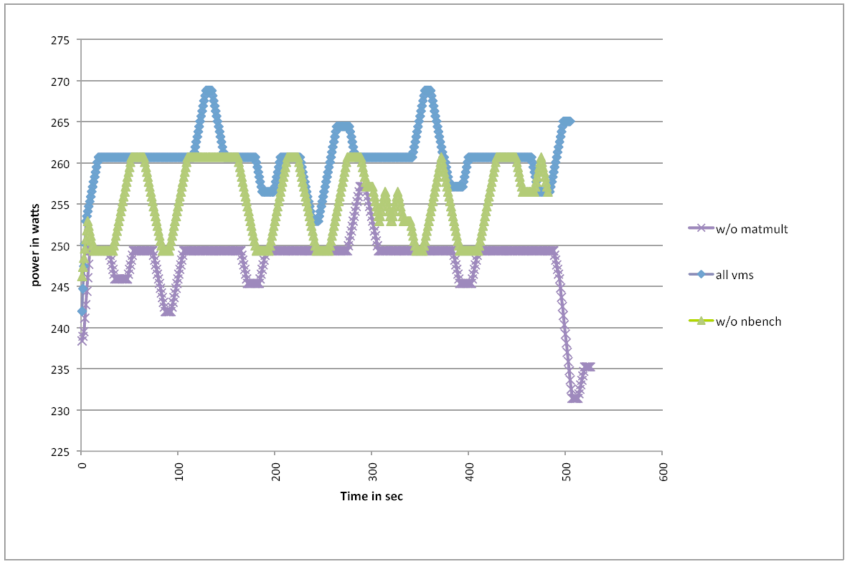
- Closing the loop with power caps and distributions derived from CEETHERM thermal models

Power-centric Load Management

- Policy:
  - Balanced power usage
  - Improve energy efficiency
    - Run all servers at reduced load vs. half of them with consolidated load?
  - Cooling considerations
    - Minimize PUE
- Consideration of heterogeneity
- Impact of reconfiguration
  - Performance perturbation and overall performance degradation
- …
- Which nodes and which VMs?

VM-level Power Metering

- Assess power and energy utilization of a VM, or a VM ensemble
- Use information in power-centric management policies
  - e.g., minimize number of VMs to migrated to reach power cap
- Use information in power-centric 'billing' policies
  - e.g., charge-back algorithm to translate power into CPU, memory, I/O resources, as needed…

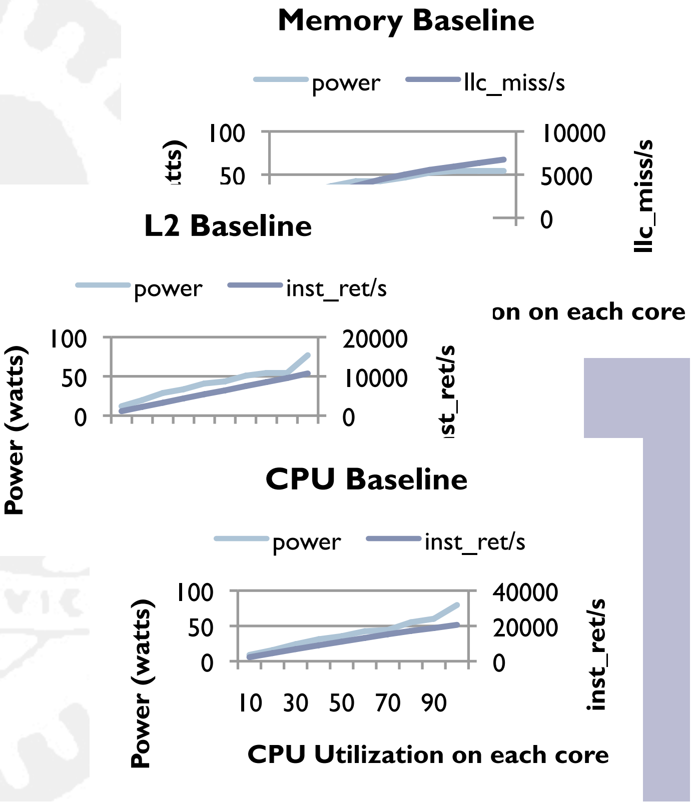


VM-level Power Metering: Our Approach

- Built power profiles for various platform resources
  - CPU, memory, cache, I/O…
- Utilize low-level hardware counters to track resource utilization on per VM basis
  - xenoprofile, IPMI, Xen tools…
  - track sets of VMs separately
- Apply monitored information to power model to determine VM power utilization at runtime
  - in contrast to static purely profile-based approaches

VM Power Model

- Resource power modeling methodology
  - determine static power component
  - wileE benchmark (SPEC Power)
  - use hardware counters and Xen monitoring tools
- Initial consideration on CPU, LLC, Memory
  - Network I/O low overall contribution
- VM predicted power model:



Predicted power = nbench VMs inst_ret/s * cpu watts/inst
+ matrix mult VMs inst_ret/s * L2 watts/inst
+ matrix mul VMs llc_miss/s * Mem watts/llc_miss

# Nbench VMs	# Matrix VMs	Total Measured Power	Total Predicted Power
3	3	85.046	85.36
2	2	57.67	54.36

Easy… right… ?!

- Moving to a dual-socket quad-core platform
  - Consideration of core-socket mappings
  - FSB saturation – non-linear memory model
  - snooping traffic – significant cause of possible overestimation
    - For mixes of CPU bound VMs model very accurate
    - Once memory bound VM included – significant error – up to 25.9W for a mix of 7CPU + 1Mm bound VM!
- Moving to a Nehalem platform
  - Inclusive caches – accuracy of existing model improved with Mm bound VMs too.
  - Ah… NUMA! … start with mix with single Mm-bound VM first…
  - 2 CPUs < 1 CPU?

benchmark	System power	Dynamic power	Predicted power	error
povray	225.8	51.8	51.17	0.63
namd	225.1	51.1	50.02	1.08
Lbm	230.2	56.2	57.06	-0.86
gobmk	226.1	52.1	48.31	3.78
h264ref	225.8	51.8	51.72	0.08

Ongoing work

- Continuing to try to make sense of it all! – Understand feasibility, utility and limitations of the approach
- Important observation:
  - How power utilization is assessed is a platform property!
    - Approaches based on application profiles will have limited applicability
    - Same for approaches which ignore interactions with the memory subsystem
  - Dynamic monitoring adds overhead, but acceptable

- Apply to distributed management policy
  - VPMTokens
  - Energy-based charge back resource management algorithms

Monitoring overhead	w/o monitoring	Monitoring	mon 5s sleep
nbench	1010	1022	1013
bzip2	747	854	756
milc	954	1030	964
h264ref	1090	1180	1100