TECHNISCHE UNIVERSITÄT DRESDEN
Center for Information Services and High Performance Computing (ZIH)

Node Power Consumption Variability

Energy Efficient HPC WG Workshop, November 17th 2014

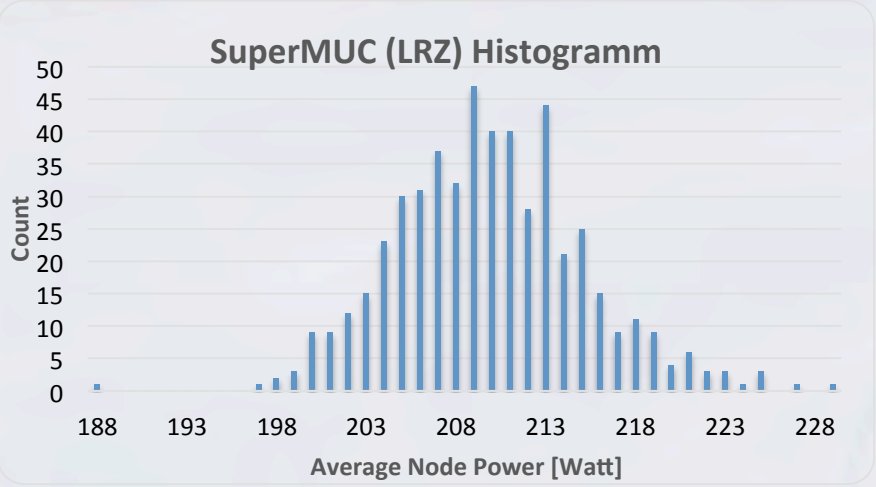
Daniel Hackenberg (daniel.hackenberg@tu-dresden.de)

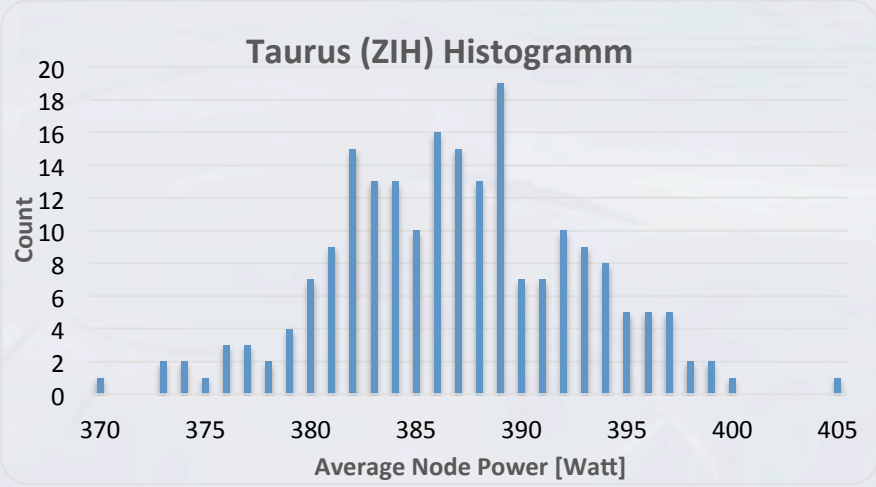
ZIH
Center for Information Services & High Performance Computing

Motivation and Challenges to Create Large Scale HPC Energy Efficiency Metrics

- Methodology needed to compare different systems at different sites
- Existing, excellent methodology from SPEC requires calibrated power analyzers for full system
- Find good sweet spot for accuracy and complexity
  - Be accurate enough to be trustworthy
  - Avoid complexity in terms of setup, measurement devices etc.
- Identify sources of variations
- Identify workload (for now: HPL and GFLOPS/W)
- Variation #1: physical between nodes
  - Measuring different nodes using
  - Identical per-node workloads
- Variation #2: logical between (MPI) ranks
  - Measuring identical nodes using a
  - Parallel workload
- Variation #3: temporal
  - Power consumption variability over time
- For HPL, the workload is highly homogeneous (uncertainty #2 irrelevant)

Variation #1 (physical between nodes)





	SuperMUC (LRZ)	Taurus (TU Dresden)
Benchmark	Prime	FIRESTARTER
Min/Max/Avg	188/229/210	370/405/387
abs. Diff.	**41**	**35**
rel. Diff.	**19,5%**	**9,0%**
CPU	2x Intel E5-2680	2x Intel E5-2690
RAM	32 GB	32 GB

Variation #1 (physical between nodes): Reducing the Sample Size

- Chernoff-Hoeffding bound calculations by Suzanne Rivoire, Sonoma State University

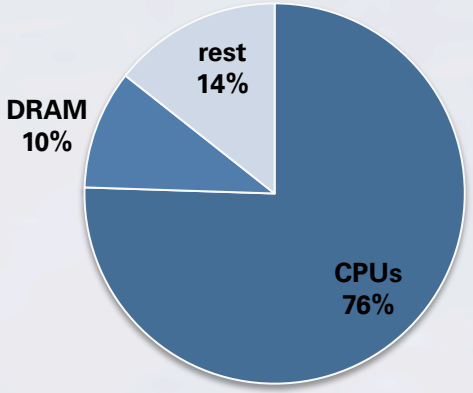
$$q \geq \frac{-I^2 \ln(p/2)}{2\delta^2}$$

allowable per-node estimation error δ	δ = 1.5%		δ = 5%	
desired probability p for estimation error to be greater than δ	1%	5%	1%	5%
Prime on SuperMUC, I = 32W	274	191	25	18
Prime on SuperMUC, I = 41W	449	313	42	29
FIRESTARTER on taurus, I = 35W	97	68	9	7

Variation #1 (physical between nodes): Causes and Developments

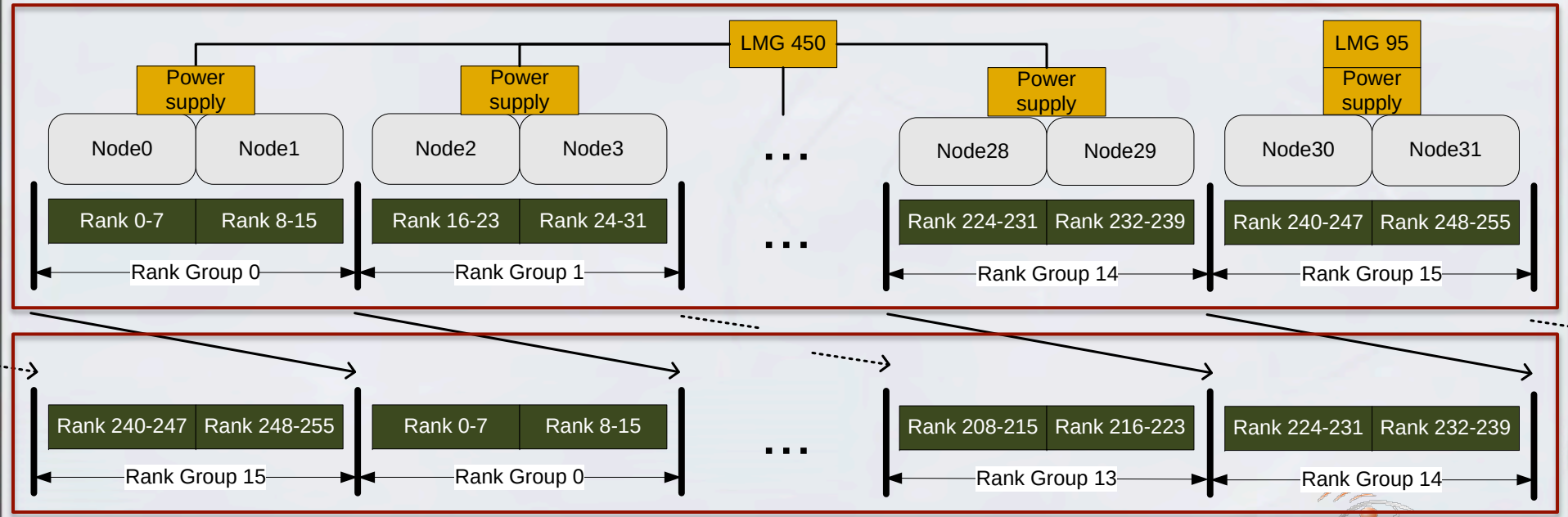
- CPUs are the dominant power consumer in HPC nodes
- Continuing trend towards integration will further increase CPU fraction of node power
- Consequently, CPU power variations are most important
- CPU power variations driven by
  - Variations in the manufacturing process
  - Varying temperature throughout the system
- Sophisticated power control units (PCUs) may change the game a little:
  - Less power variations
  - More performance variations

Power consumption breakdown for FIRESTARTER on taurus



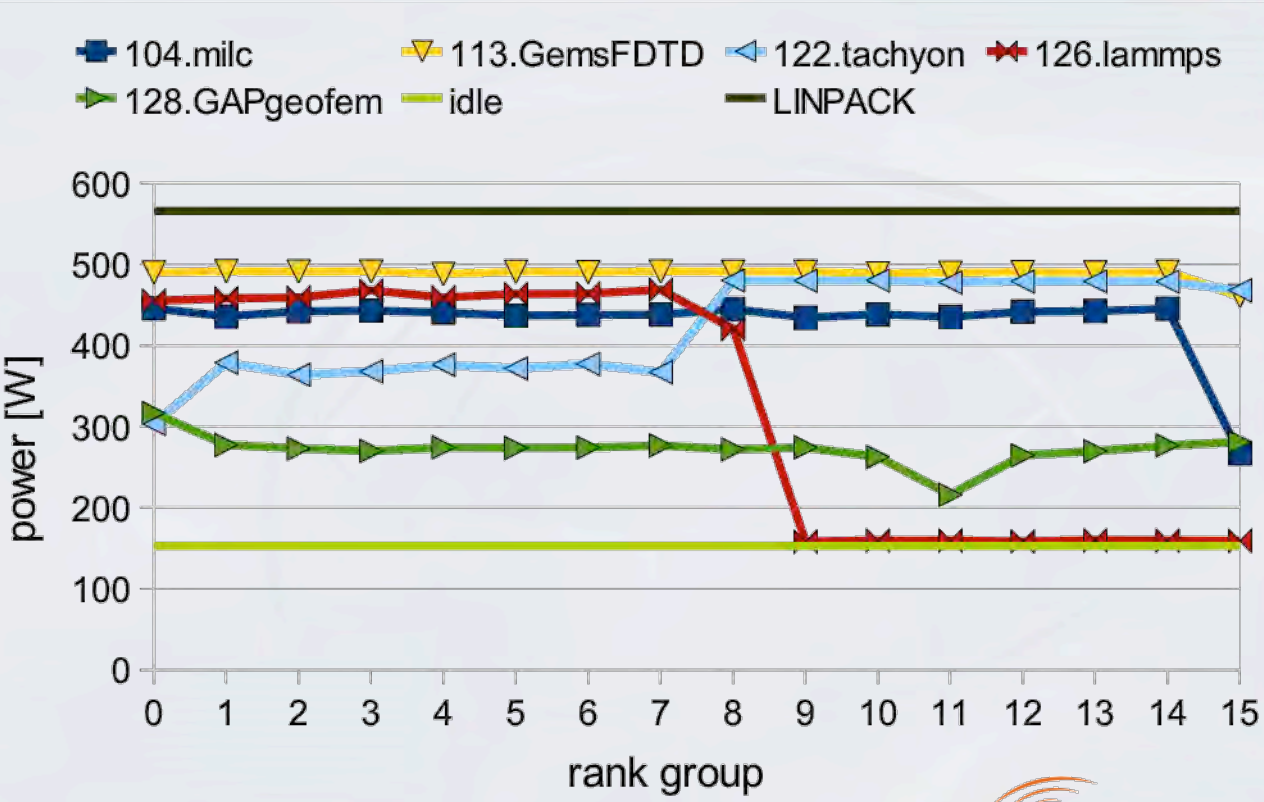
Variation #2: Logical Between Ranks

- Test setup with 16 MPI rank groups, each group has 16 MPI ranks
- MPI rank groups cycle through 16 double-nodes



Variation #2: Logical Between Ranks (2)

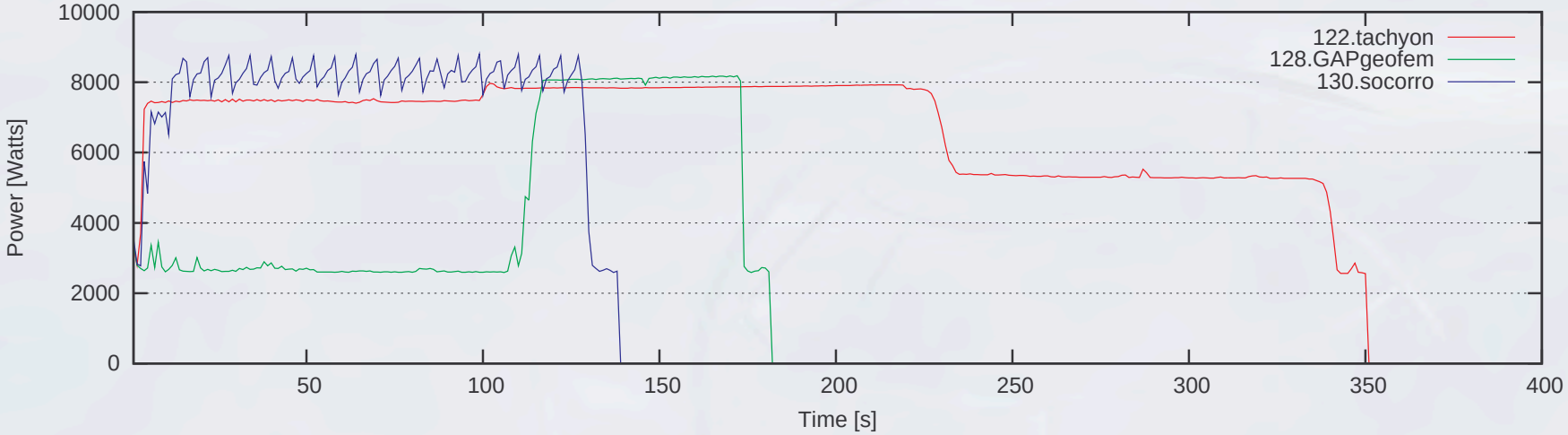
- 4-5 SPEC MPI2007 benchmarks show significant power variations
- There is no single MPI rank group that can be used for a good extrapolation
- For SPEC MPI2007, using the first rank group(s) usually works (you do not underestimate, except for tachyon)



D. Hackenberg et.al., Quantifying power consumption variations of HPC systems using SPEC MPI benchmarks, EnA-HPC 2010

Variation #3: Temporal

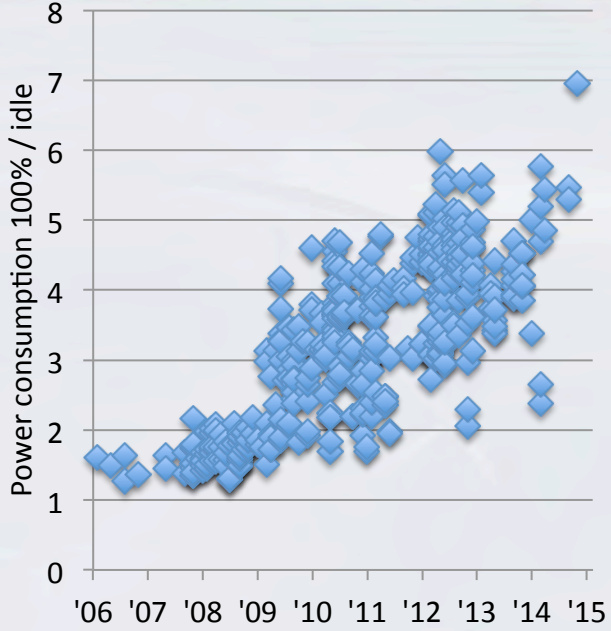
- Power consumption over time for three SPEC MPI 2007 benchmarks



- HPL also has non-constant power consumption (see presentation by Tom Scogland)
  - Initialization, computation, verification
  - Even for computation, power tail-off gets longer (Blue Gene/Q) or much longer (GPU accelerated)

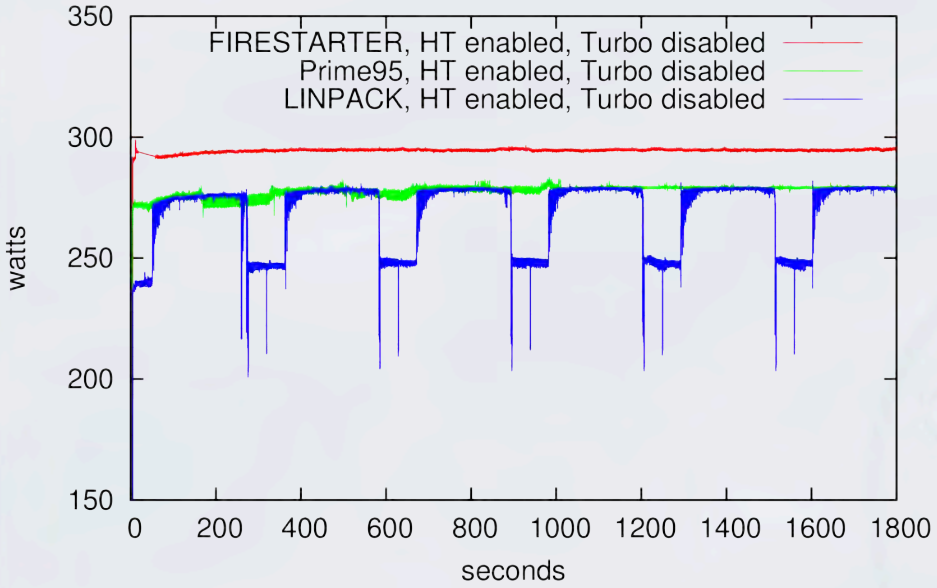
Summary and Outlook

- Three causes for variations: #1 physical between nodes, #2 logical between ranks, #3 temporal
- Need to avoid impact of variations on system metrics, preferably without doing full-system-measurements
- Options to tackle #1 and #3 are being evaluated
- #2 needs to be considered, maybe even for HPL
- P(full_load)/P(idle) is steadily increasing
  - This increases power variations #2 (logical between ranks) and #3 (temporal)
- PCUs may decrease power variations and increase performance variations

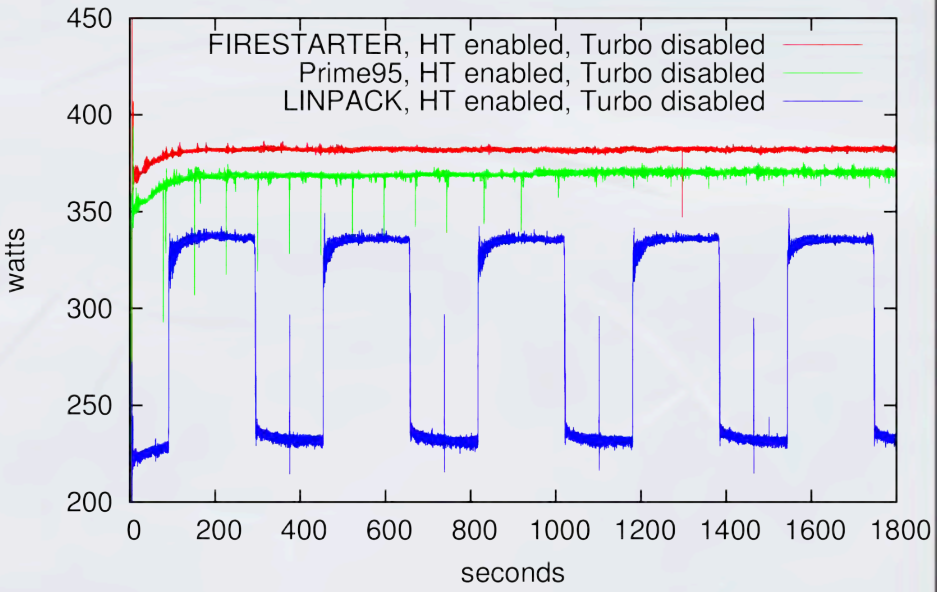


- SPECpower Benchmark
- Full load vs. idle

FIRESTARTER: A Processor Stress Test Utility



Intel Xeon X5670, Westmere-EP (2P), SSE routine



Intel Xeon E5-2670, Sandy Bridge-EP (2P), AVX routine

http://tu-dresden.de/zih/firestarter/