TECHNISCHE UNIVERSITÄT DRESDEN

Center for Information Services and High Performance Computing (ZIH)

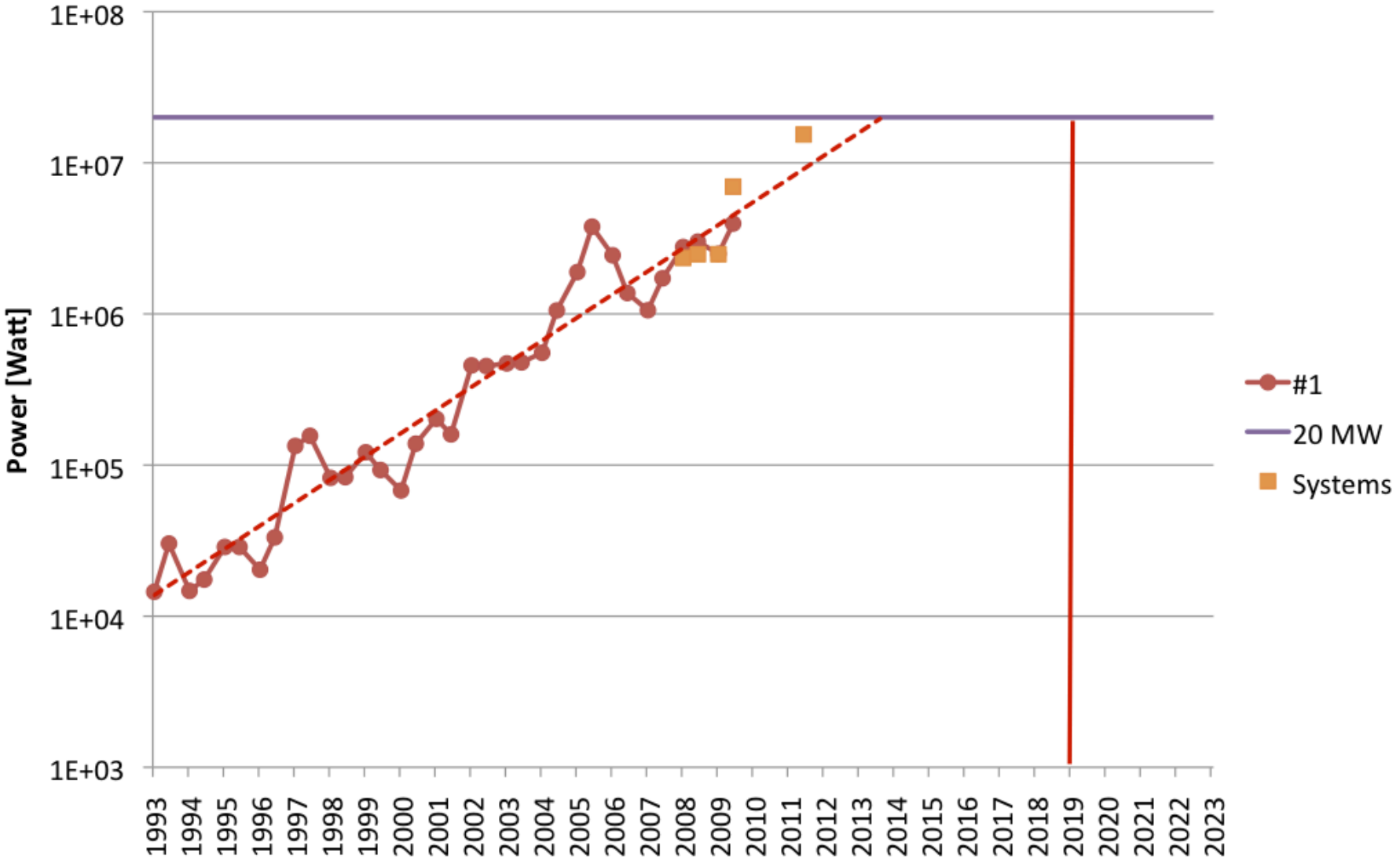
Quantifying power consumption variations of HPC systems using SPEC MPI benchmarks

EnA-HPC, Sept 16th 2010

Daniel Hackenberg, Robert Schöne, Daniel Molka,
Matthias S. Müller, Andreas Knüpfer

(mailto: daniel.hackenberg@tu-dresden.de)

Motivation: Power Consumption Extrapolation



HPC Power Consumption Benchmarks

- The HPC community needs power consumption benchmarks!
- How about the Green500 list?
- TU Dresden LNXI Cluster reported at June 2008 list:
  - 2576 AMD Opteron cores, 2.6 GHz
  - Green500 power consumption: 206 kW (peak)
  - Actual power consumption: ~300 kW

What we have: The Green500 List

Run Rules

Run rules for measuring the power and energy consumption of qualified high-end clusters: We will be updating and reposting this information soon. For more information email info AT green500.org

Tutorials

Power Measurement Tutorial for the Green500 List [pdf]

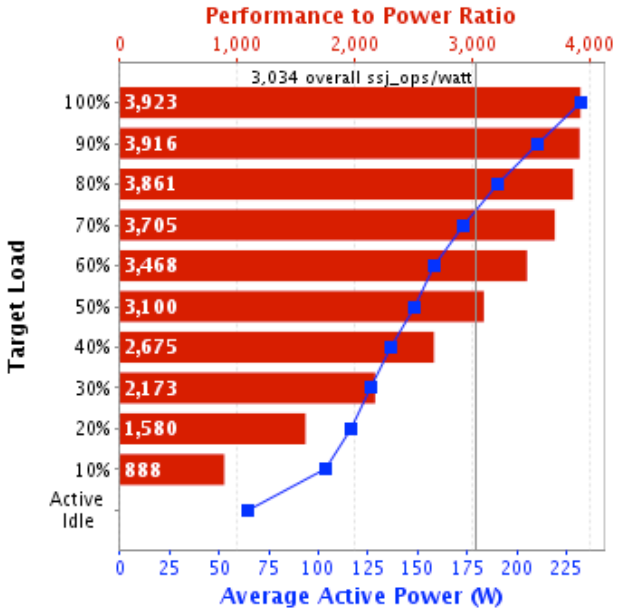
R. Ge, X. Feng, H. Pyla, K. Cameron, and W. Feng.

Source: green500.org

- **Power extrapolation: $P = N \cdot P_{unit}$**
- **Assuming that**
  - **(1) the computational workload during the Linpack benchmark is well-balanced across all units, and**
  - **(2) all units are identical and consume the same amount of power for the same workload.**

What we have: SPECpower_ssj2008

- Industry-standard benchmark that
  - Evaluates the power and performance characteristics of
  - Volume server class and multi-node class computers
- Very mature benchmark methodology
  - Dedicated power daemon for accepted power meters only
  - Records current, voltage, power factor, temperature
  - Very strict run rules
- BUT
  - Java Workload
  - Strongly VM dependent
  - Servers typically run Windows



What We Need

- An HPC benchmark with a sophisticated power measurement methodology

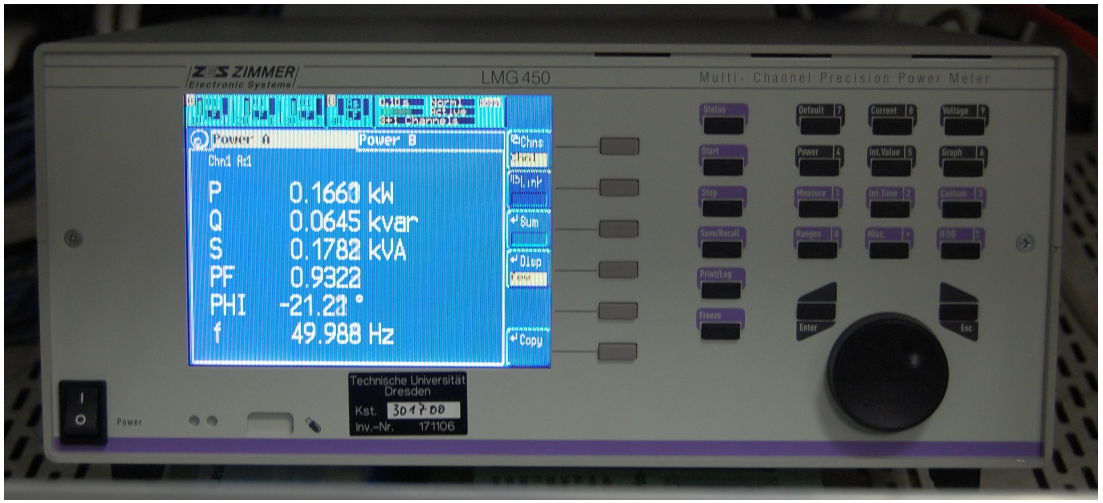
- SPEC MPI2007
  - Industry-standard HPC benchmark
  - Averages the results of 12 MPI applications that are common in HPC
  - Medium data set scales up to 128 MPI ranks, runs up to 512 ranks
  - Large data set scales up to 2048 ranks, tested up to 4096 ranks

- Challenge:
  - Companies probably do not want to measure each node
  - We need run rules that ensure that hardware vendors do not cheat (too much)

Test Equipment Overview

- Test System: IBM iDataPlex
  - 32 nodes, each with 12x 4 GB DDR3, 250 GB HDD, QDR IB
  - 64x Intel Xeon E5530, 2.4 GHz, 80W TDP
  - Turbo Boost up to 2.66 GHz, HyperThreading off
- Power Meter
  - 1x ZES LMG 450 (4 channel)
  - 1x ZES LMG 95 (1 channel)
  - Measuring compute nodes only, no switches, no I/O
- Benchmark: SPEC MPI2007 V2.0
  - medium data set
  - Intel Compiler Suite 11.1, Open MPI 1.4.1

Test System & Power Measurements



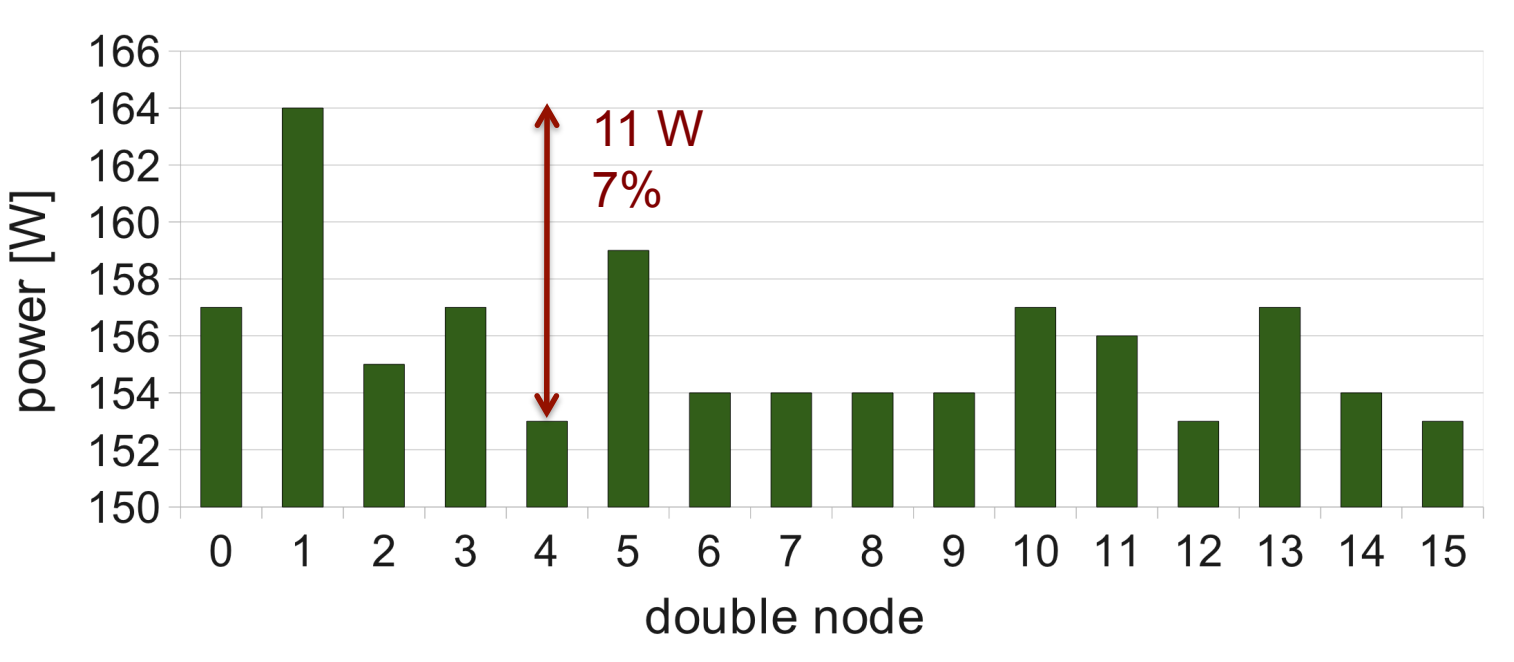
Test System & Power Measurements

Old vs. New

- Comparing the power consumption of
  - a 2005 LNXI dual core Opteron Cluster
  - a 2010 IBM iDataPlex quad core Nehalem Cluster

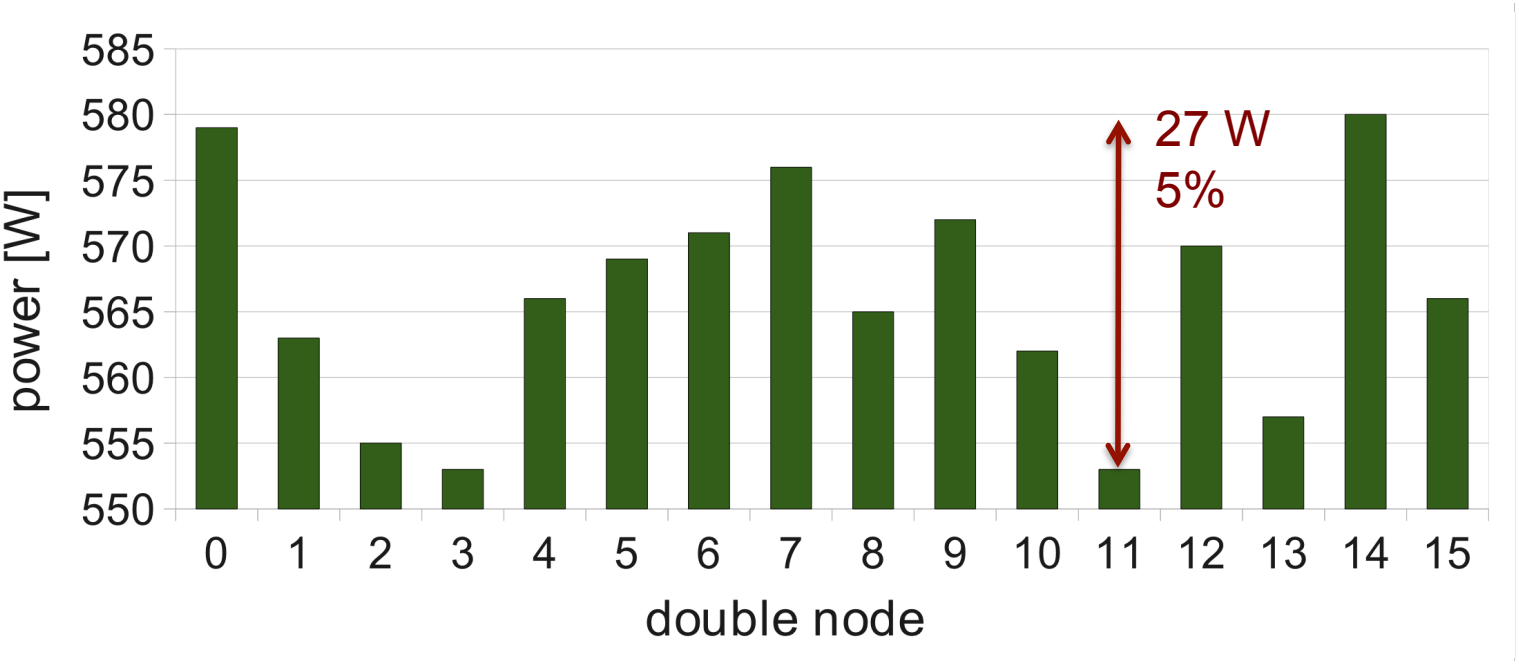
	Linpack	107.leslie3d	104.milc	Idle
Cluster 2005	100 %	92 %	87 %	65 %
Cluster 2010	100 %	83 %	71 %	26 %

Per-Node Power Consumption Variation: Idle



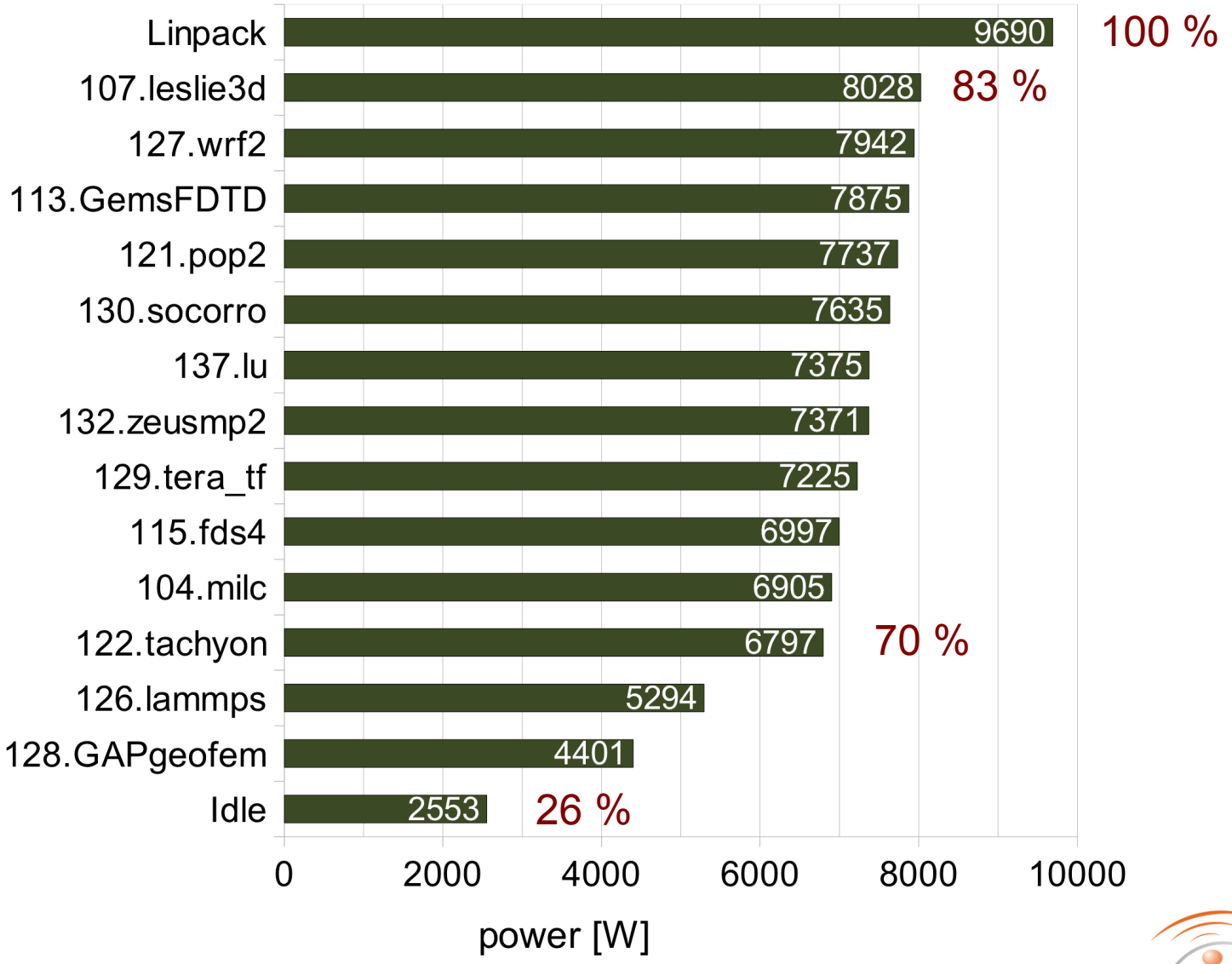
- Idle variations become bigger as soon as you swap broken parts
- We have seen >12% in our Opteron Cluster

Per-Node Power Consumption Variation: Linpack

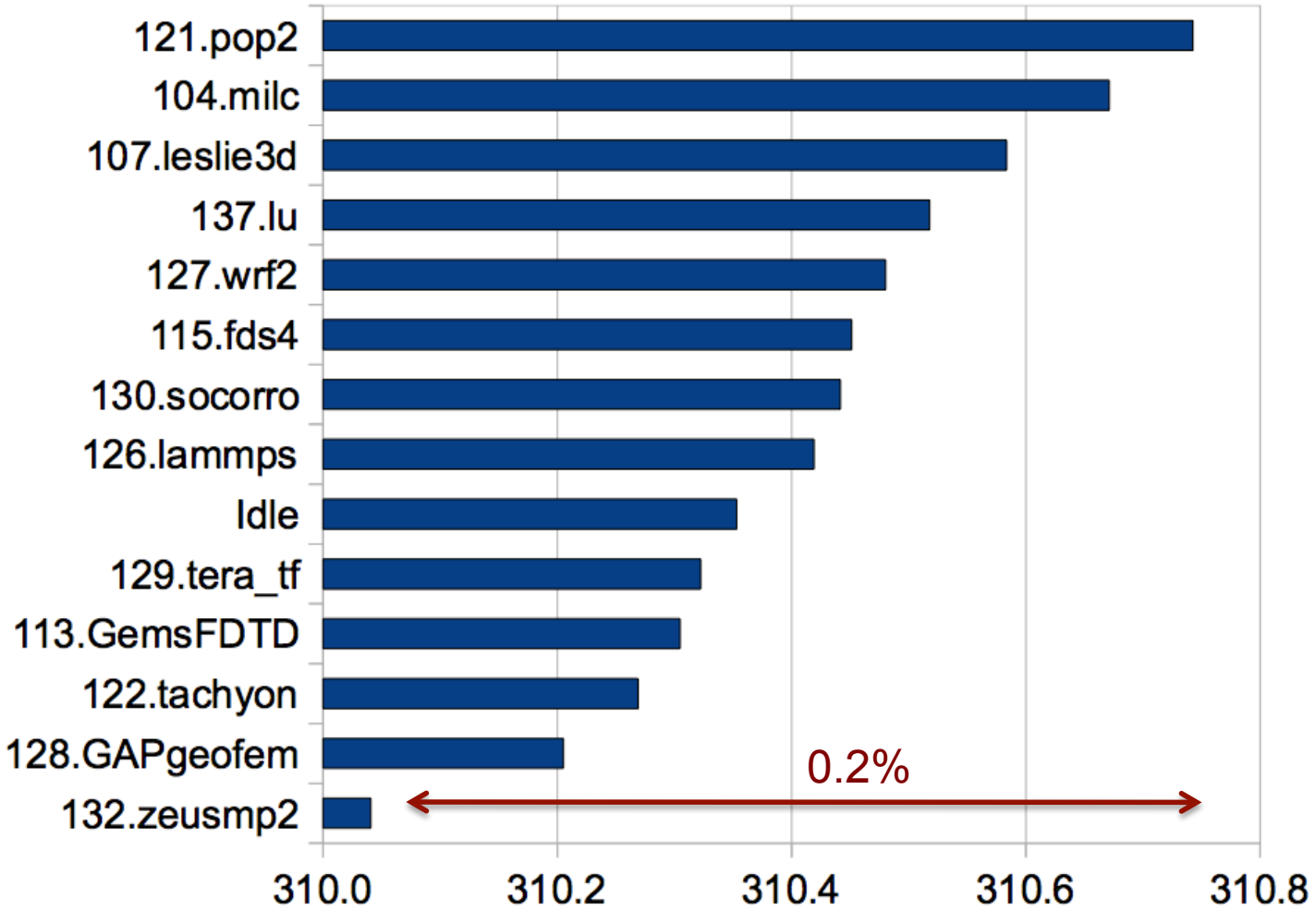


- Relative per-node power variation of idle and Linpack similar

Power Consumption Variation of 256 Intel Nehalem Cores

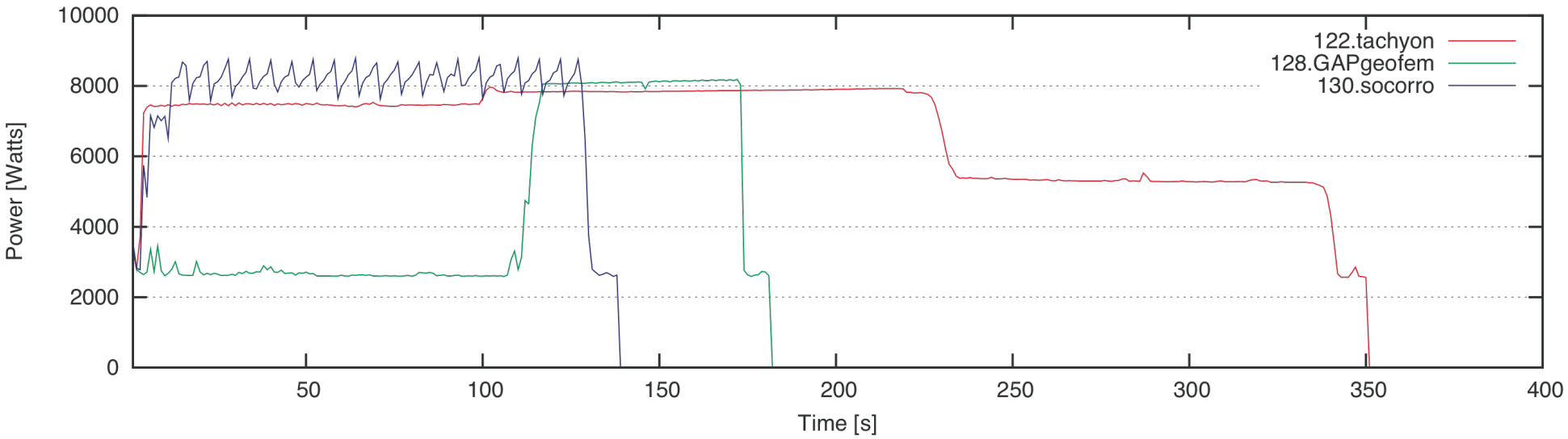


Power Consumption Variation of an IB Switch



- IB switch power consumption independent of the IB network traffic

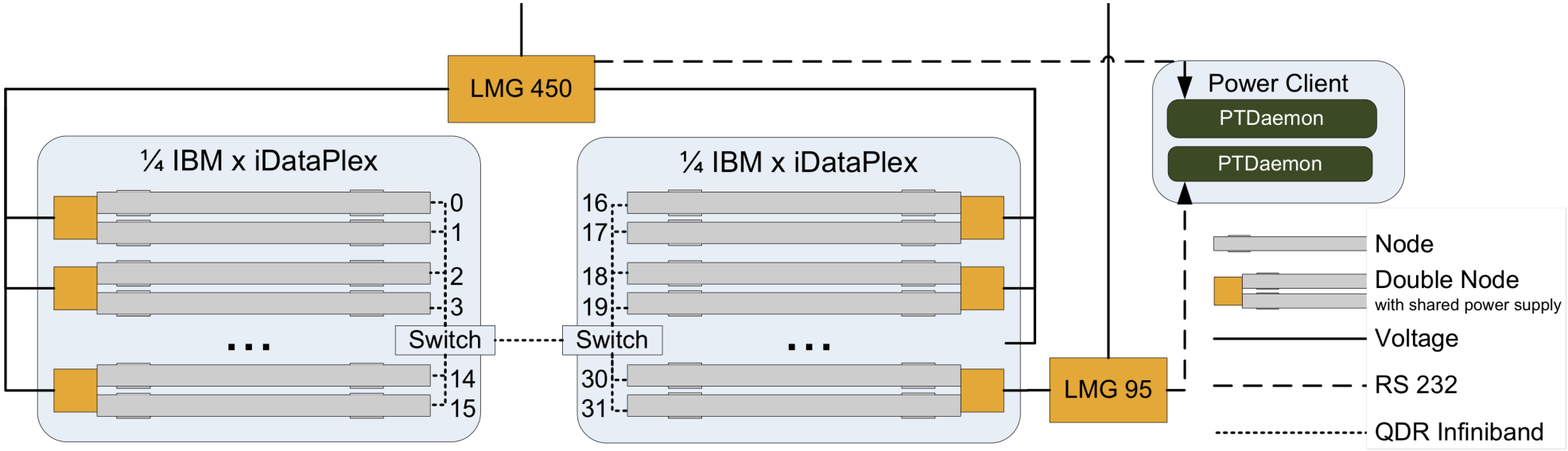
Power Consumption Variation over Runtime



- 130.socorro has a nice repetitive pattern (variation of ~1 KW)
- 128.GAPgeofem (Geophysical FEM) runs serial code for ~2/3 of the runtime
- 122.tachyon (ray tracing) has three groups of MPI ranks that differ in runtime due to the IB network setup

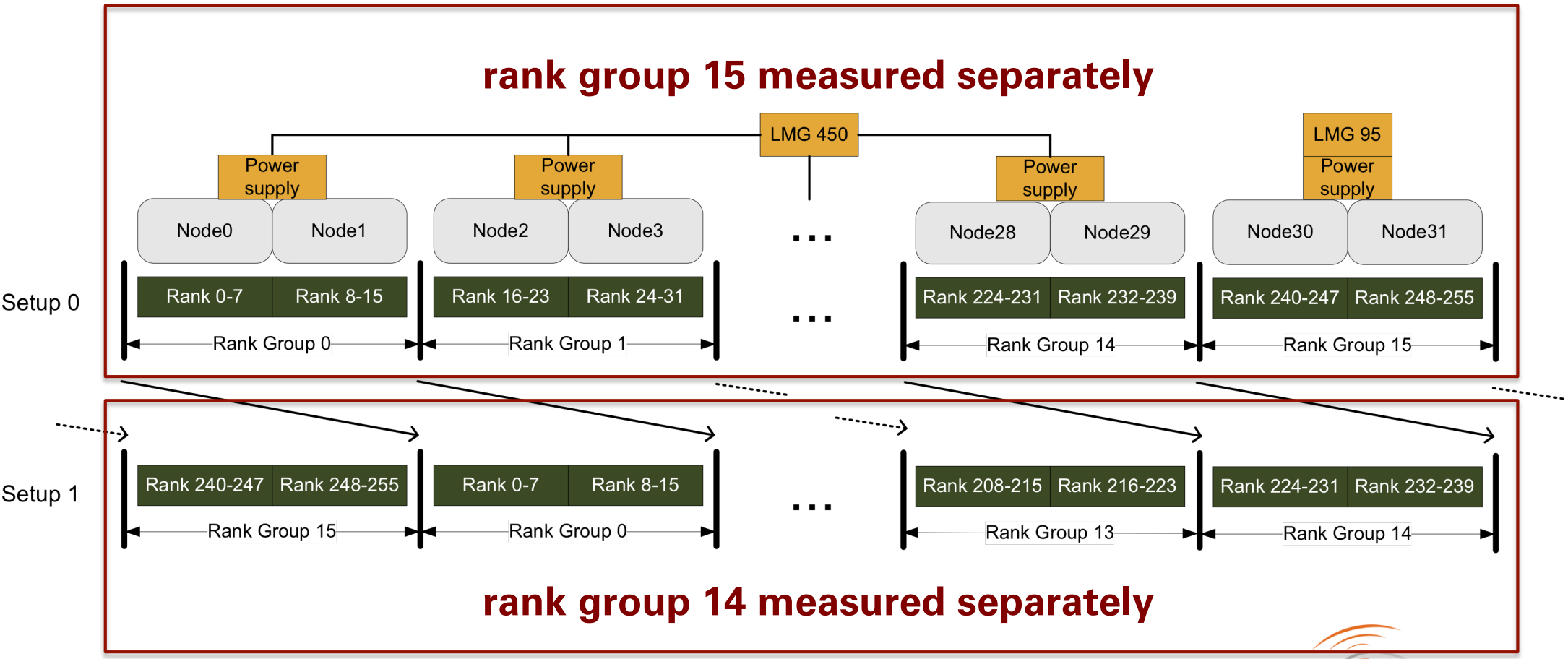
Power Measurement Software Infrastructure

- Node 0-29 (240 cores) measured by one power meter
- Node 30+31 (16 cores) measured by a second power meter
- IB switch is powered separately

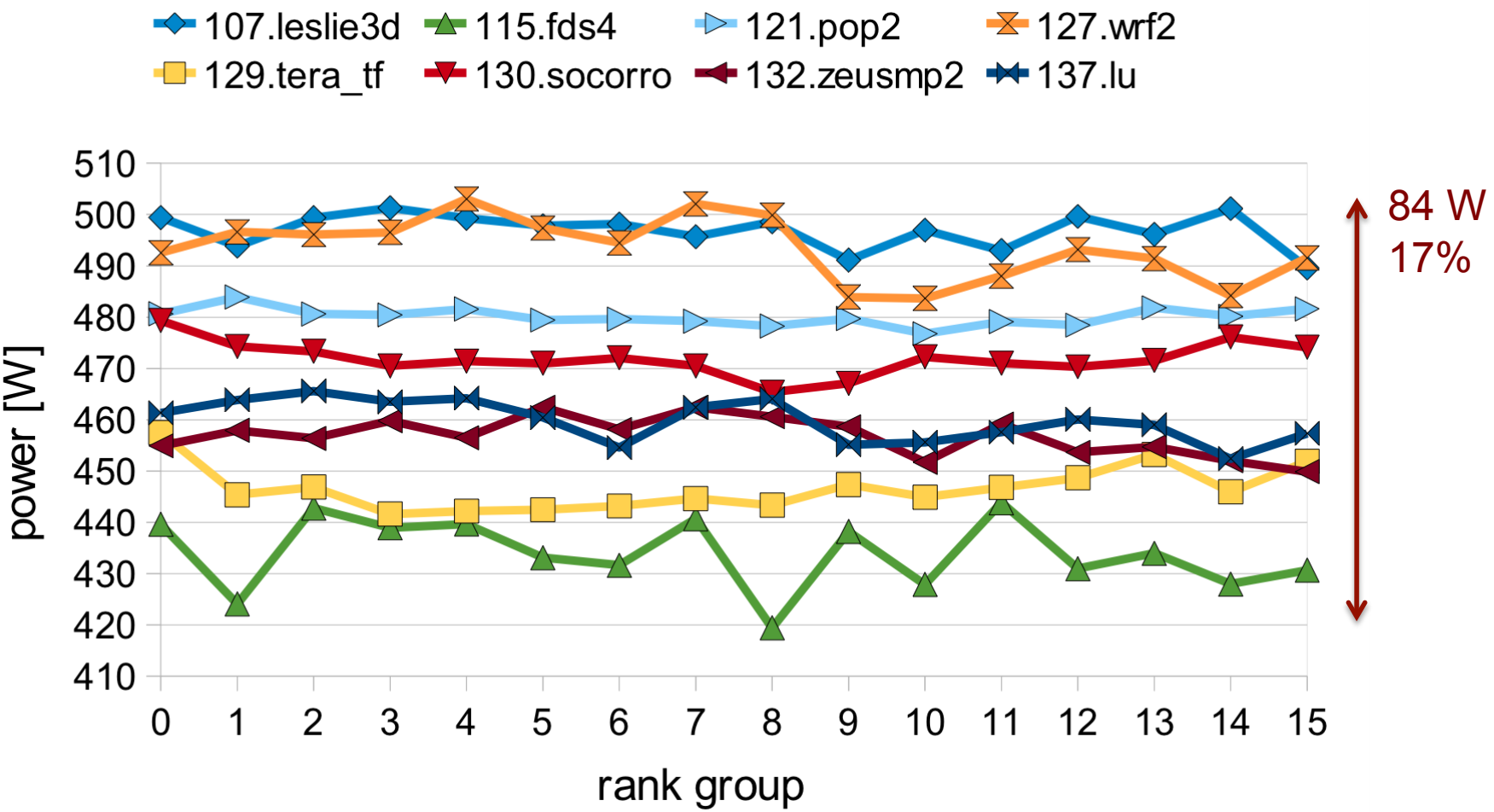


Test Series

- 16 MPI rank groups, each group has 16 MPI ranks
- MPI rank groups cycle through the 16 double-nodes

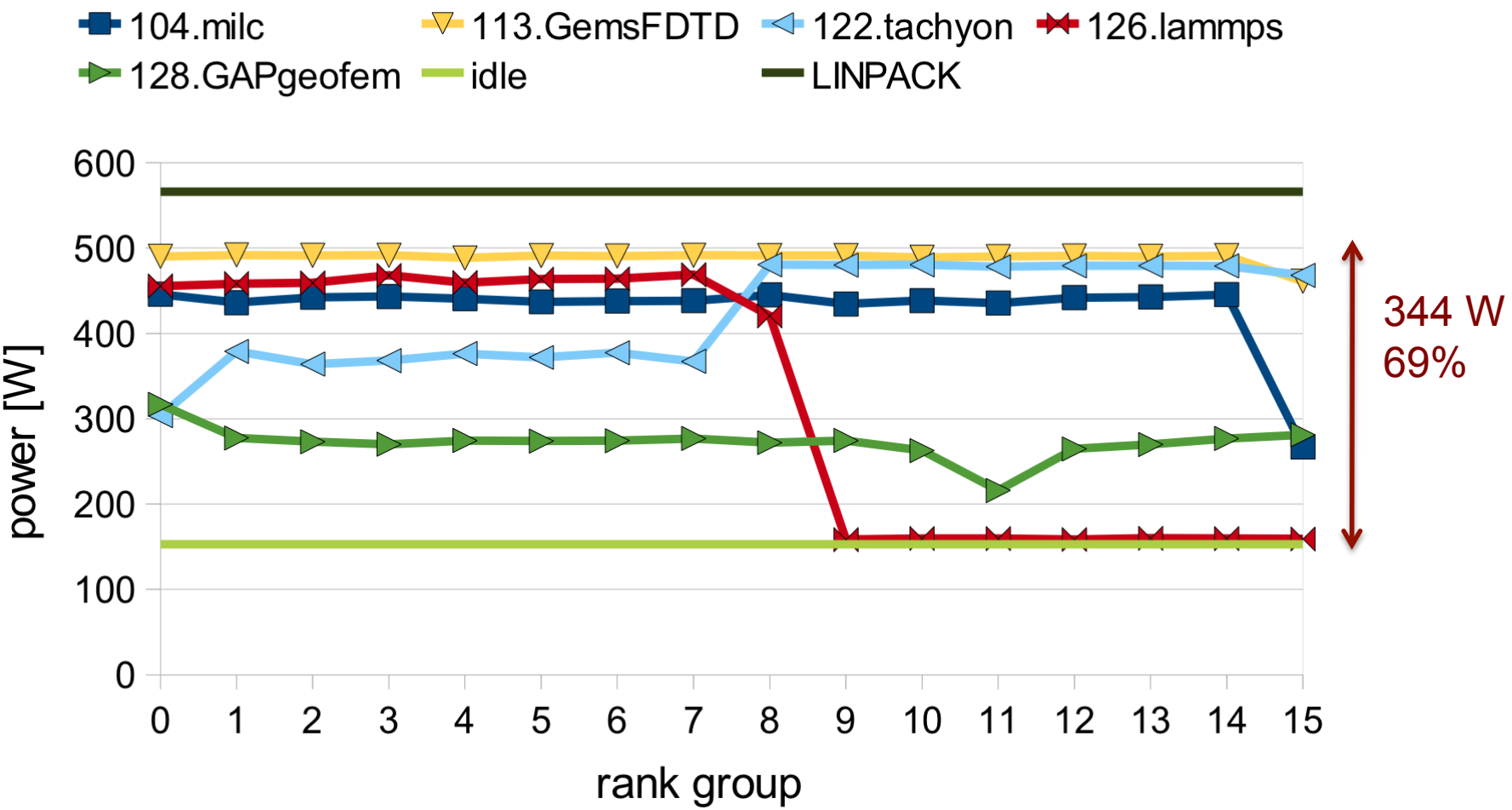


Power Consumption of MPI Rank Groups



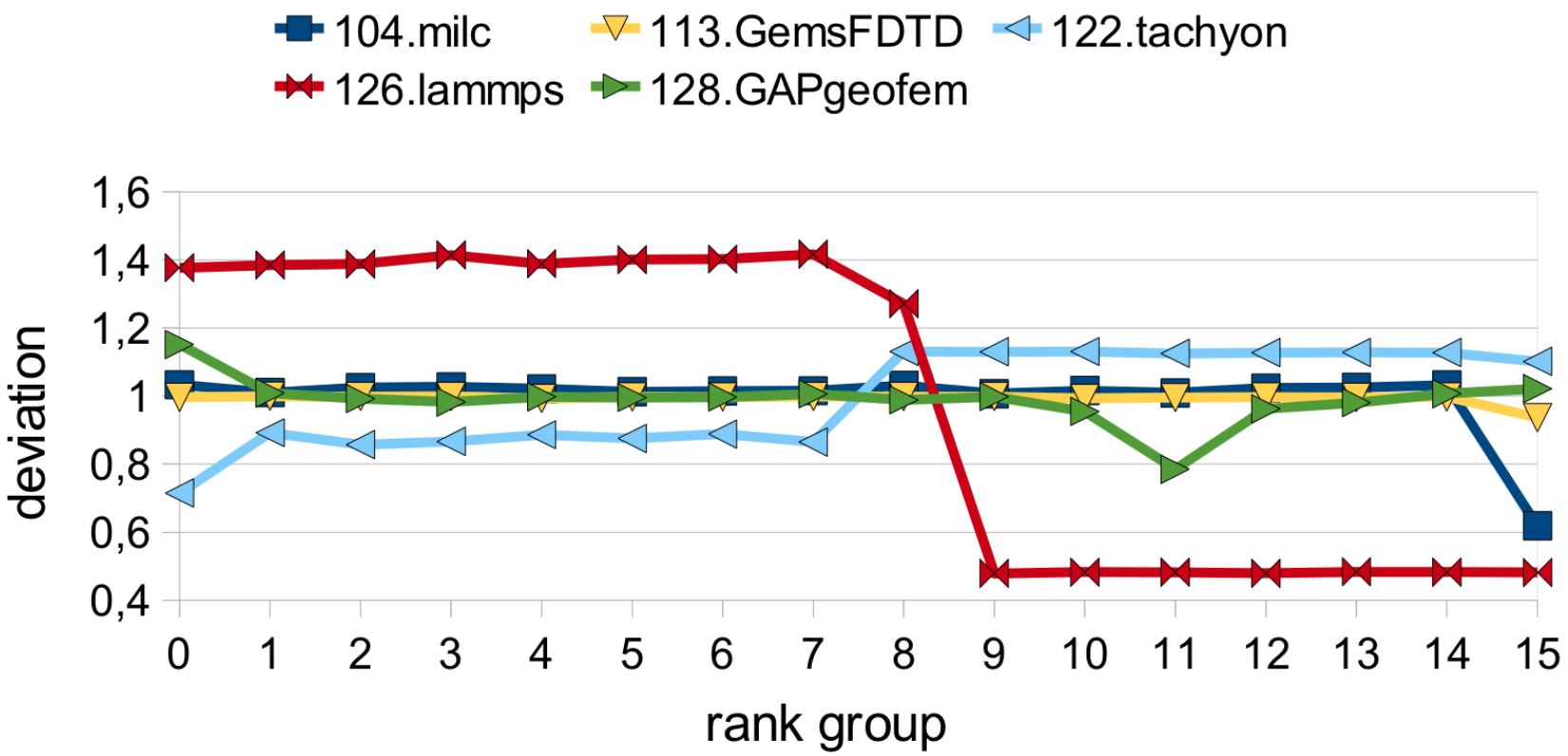
- 8 SPEC MPI2007 benchmarks show very small power variations

Power Consumption of MPI Rank Groups



- 4-5 SPEC MPI2007 benchmarks show significant power variations

Power Consumption Extrapolation Correctness



- 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)

Conclusions

- Power consumption variations are growing, therefore
- Power consumption extrapolation is difficult

- We need industry standard HPC benchmarks that
  - Include a power metric
  - Have well-defined run rules
  - Hopefully do not require companies to buy one power meter per node

- SPEC OMP/MPI2007 aims to fill that gap
- SPEC OMP currently in an early stage of defining run rules
- SPEC MPI2007 will take a bit longer

Conclusions

- Power consumption variations are growing, therefore
- Power consumption extrapolation is difficult

- We need industry standard HPC benchmarks that
  - Include a power metric
  - Have well-defined run rules
  - Hopefully do not require companies to buy one power meter per node

- SPEC OMP/MPI2007 aims to fill that gap
- SPEC OMP currently in an early stage of defining run rules
- SPEC MPI2007 will take a bit longer

Full Paper Available

- http://www.springerlink.com/content/bg6j875q82161605/