

Power Balancing in an Emulated Exascale Environment

The Twelfth Workshop on High-Performance, Power-Aware Computing (HPPAC'16)

Matthias Maiterth (LMU), Martin Schulz (LLNL),
Barry Rountree (LLNL), Dieter Kranzmüller (LMU/LRZ)

Matthias Maiterth

<http://www.nm.ifi.lmu.de/~maiterth/>



Exascale Projection (2010)

– one of Plenty



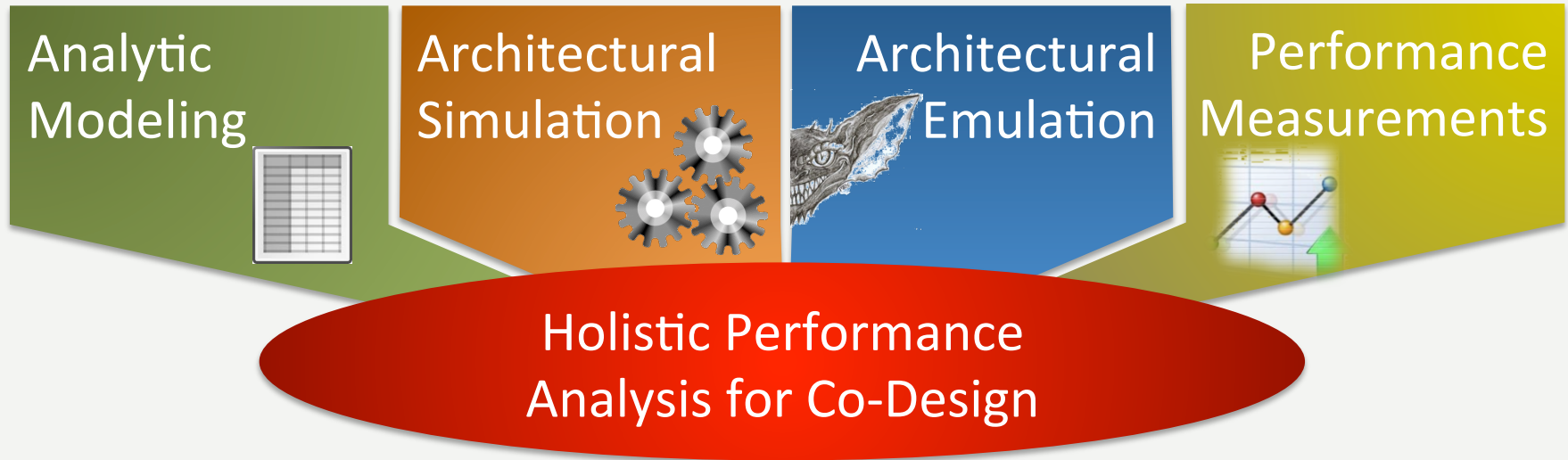
	2010	2018	Factor Change
System Peak	2Pf/s	1Ef/s	500
Power	6MW	20MW	3
System Memory	0.3PB	10PB	33
Node Performance	0.125Gf/s	10Tf/s	80
None Memory BW	25GB/s	400GB/s	16
Node Concurrency	12CPUs	1,000CPUs	83
Interconnect BW	1.5GB/s	50GB/s	33
System Size (nodes)	20K nodes	1M nodes	50
Total Concurrency	225K	1B	4,444
Storage	15PB	300PB	20
Input/Output bandwidth	0.2B/s	20TB/s	100

ASCAC Subcommittee Report: The Opportunities and Challenges of Exascale Computing,
Department of Energy, Office of Science, Fall 2010

Some thoughts:

- Horizon moved several times
- Evolutionary Approximation
- Trend-line of past trends
- Some entries seen as targets

How to use these projections?



- Emulating Exascale by Resource Restriction
- Focus on Power, Memory, Resiliency and Noise
- Scalable Framework running on current HPC systems
- Co-Design as Goal
- Make current Application and System Software Exascale-ready

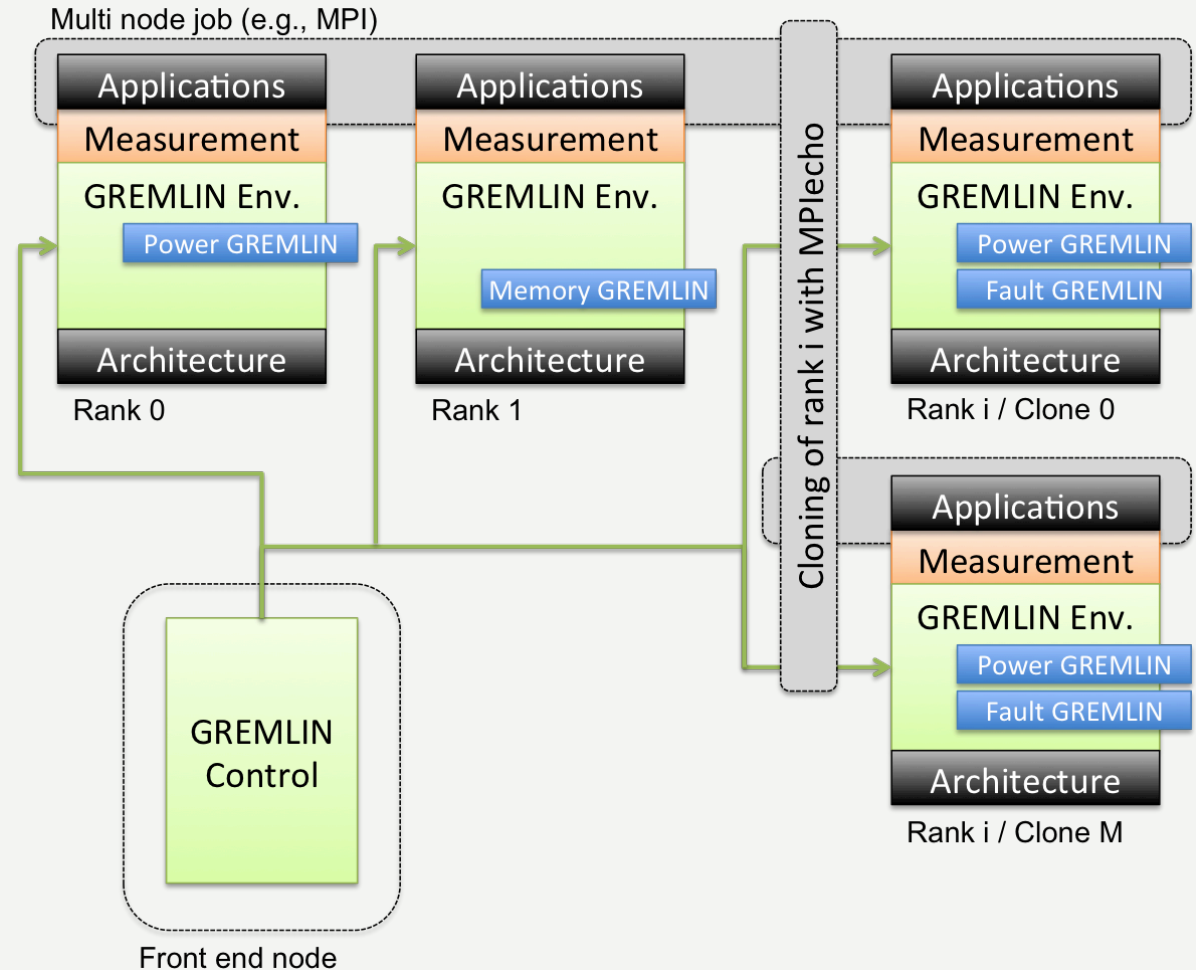
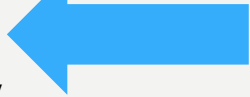


Framework for System Emulation:

- Target future platform
- Exascale

Types of GREMLINs:

- Power
- Memory
- Resiliency
- Noise

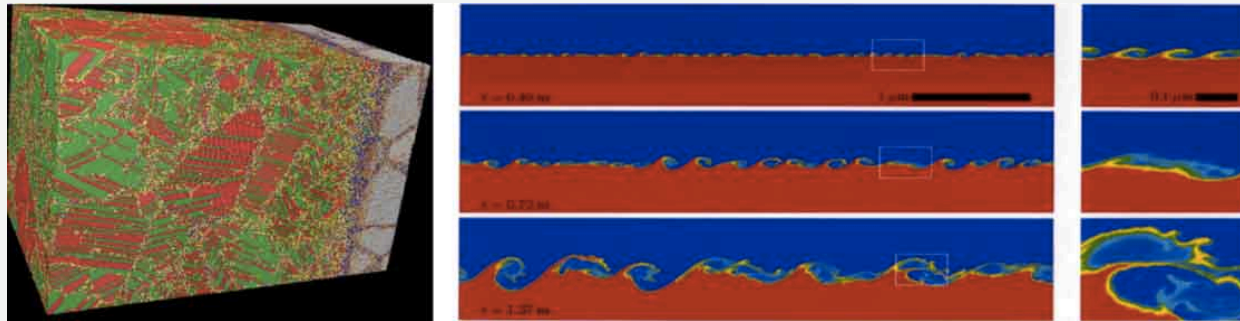




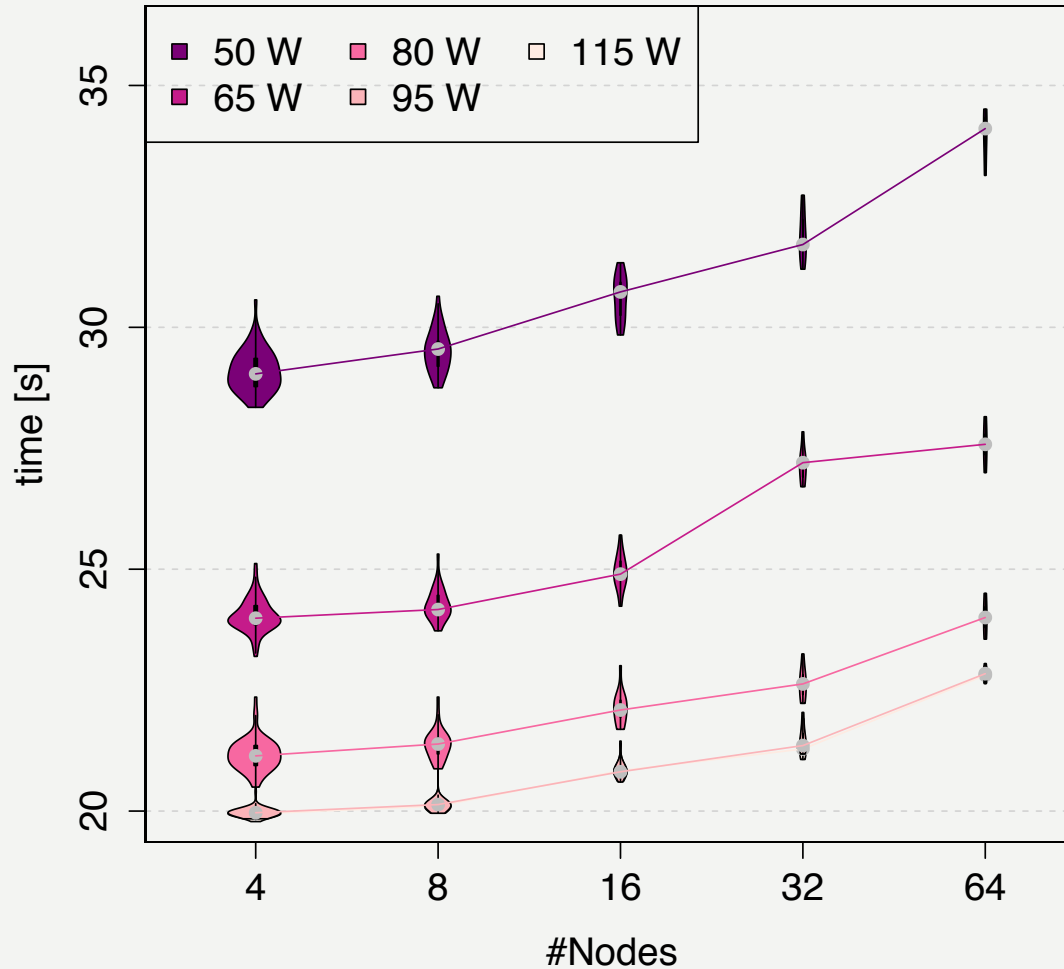
Processor architecture	Xeon 8-core E5-2670 (Intel)
Operating system	TOSS
Process clock speed	2.6 GHz
Nodes	1,296
Cores per node	16
Total cores	20,736
Memory per node	32 GB
Total memory	41.5 TB
Thermal Design Power	115 W

- Kernel module and libraries for RAPL available (libmsr & msr-safe)
- MPI implementation used: MVAPICH (ICC)

- Proxy application for classical molecular dynamics
- Developed by ExMatEx DOE Co-Design Center



- Using 16 MPI processes per Node
- Measured Power consumption: ~85W
- Run on up to 256 nodes, weakly scaled.
- Used in the following to study behavior for power limited system.

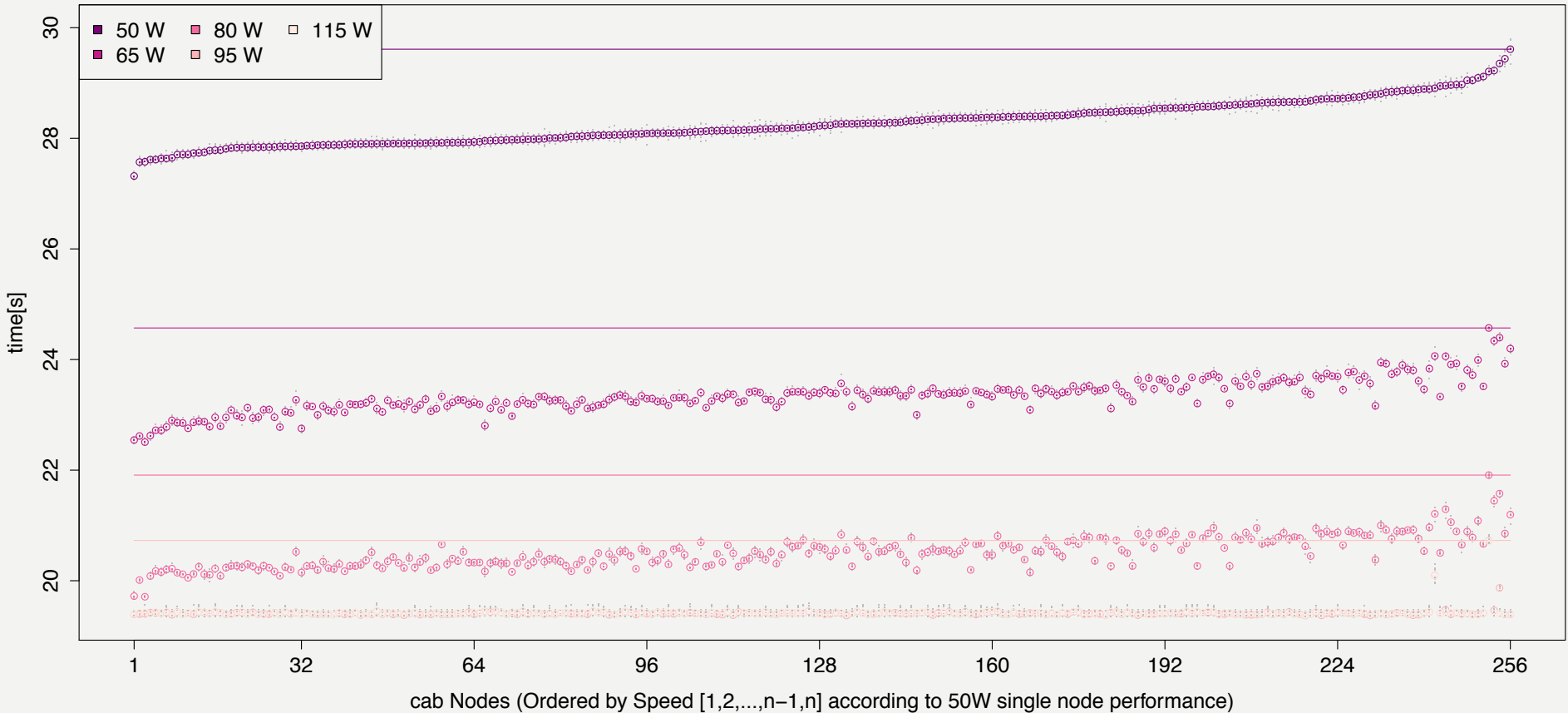


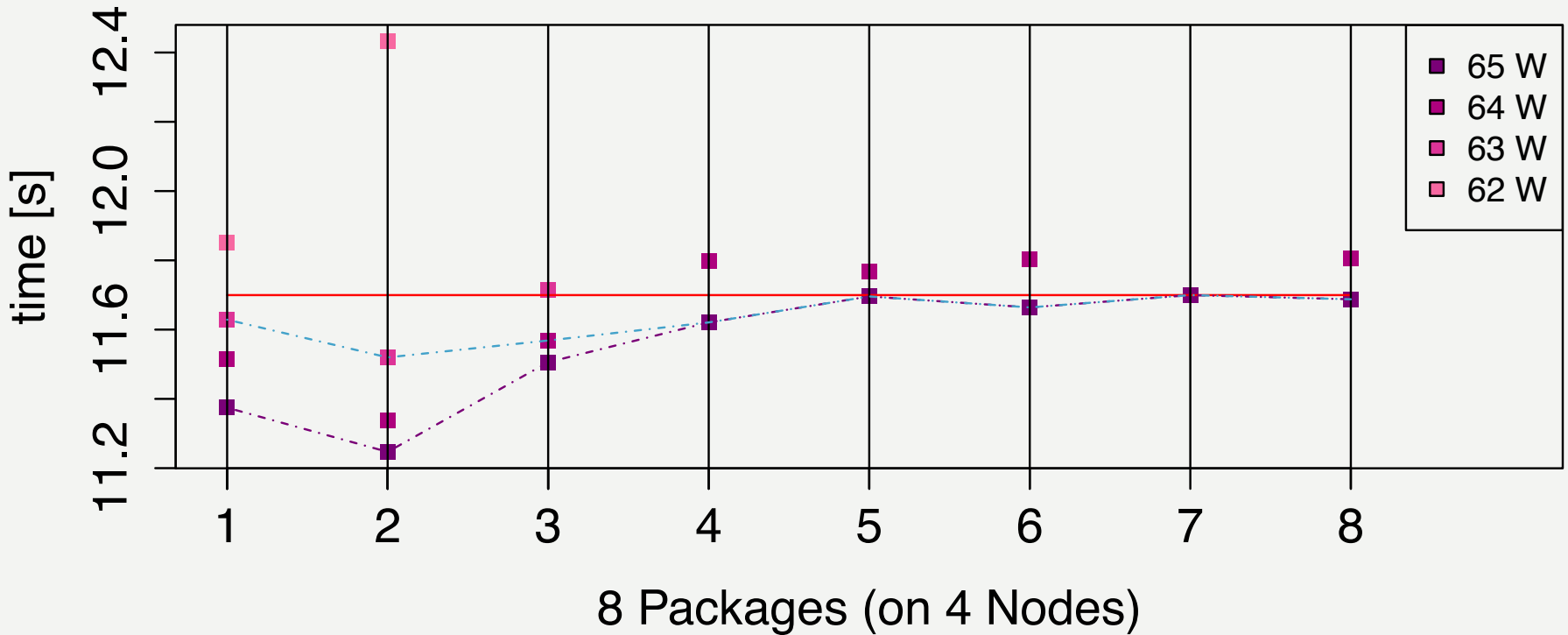
Notes on CoMD:

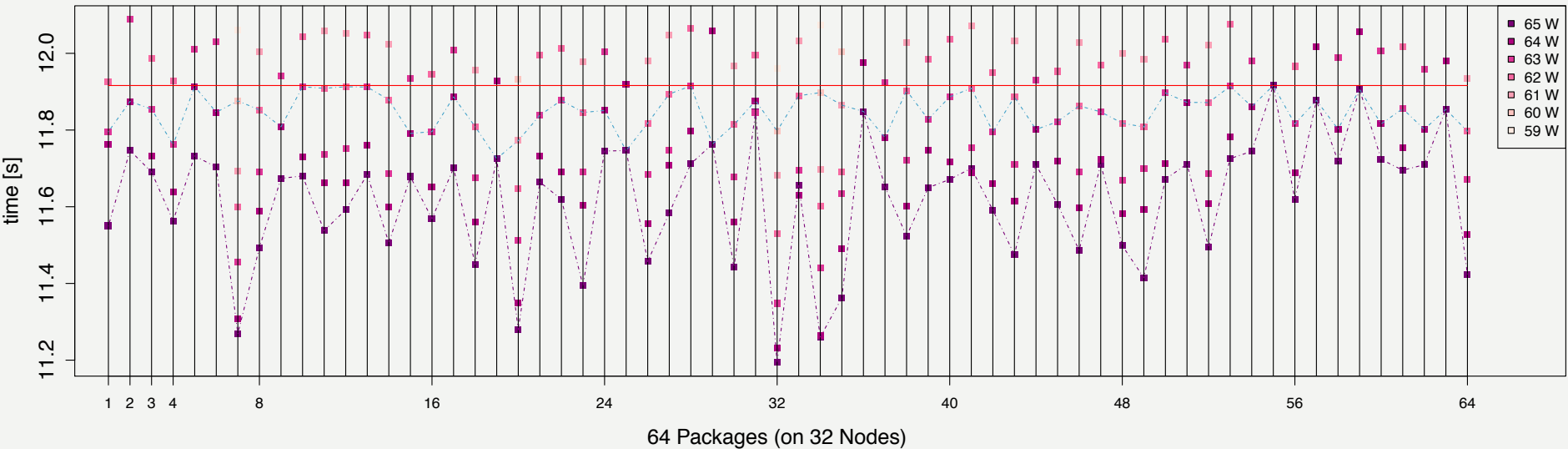
- ~85W Power draw

Notes on Plot:

- Set of 256 Nodes returned by slurm
- 256/#Nodes reps
- Average over 5 measurements



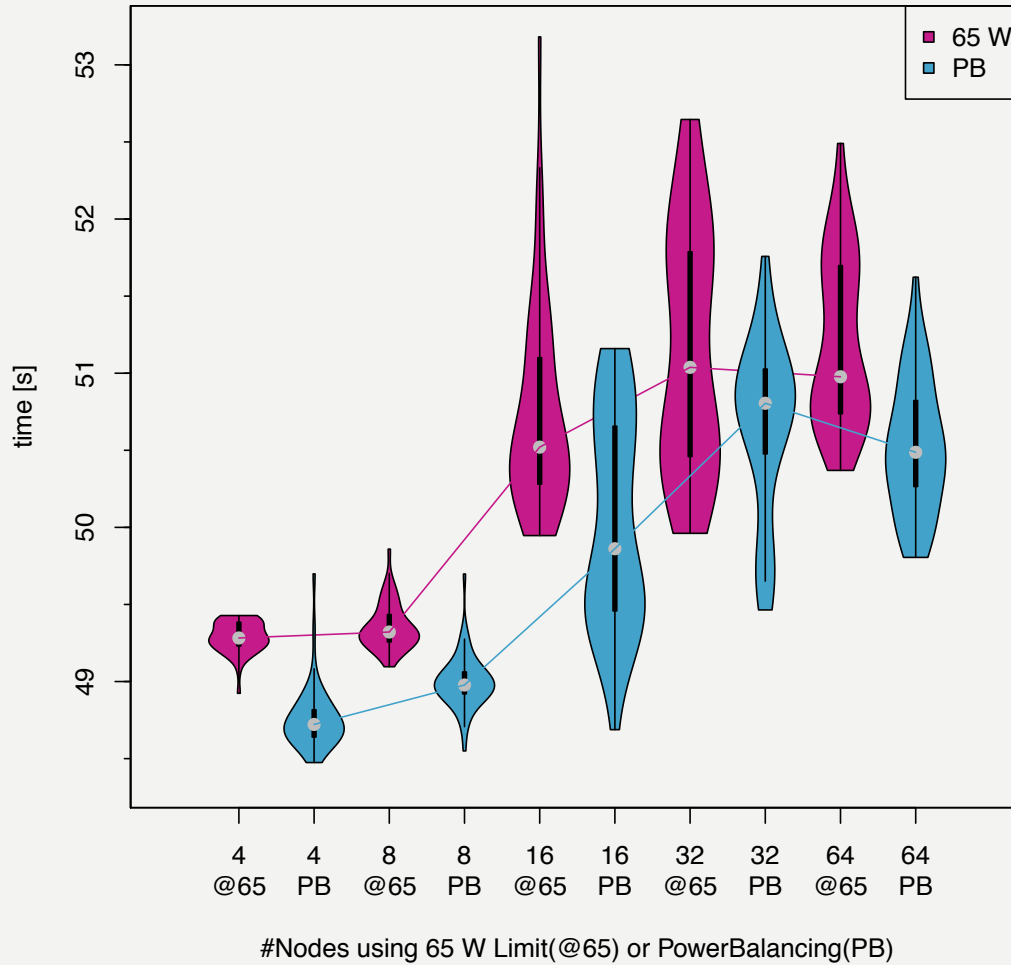




- Possible to reclaim 129W
- But: Multinode performance impacted! → Reintroducing Power

Scaling Study of CoMD

Balancing 4 to 64 nodes



Scaling Study of CoMD

Balancing 4 to 64 nodes



Nodes	Performance difference @ 65 W	Power reclaimed	Power imp. Per Package	Total Power shifted	PB Mean Speedup
4	5.75%	14W	1W	8W	1.15%
8	6.60%	23W	1W	16W	0.70%
16	6.71%	68W	2W	64W	1.32%
32	7.63%	189W	2W	128W	0.46%
64	8.05%	428W	3W	384W	0.97%



- Improving Algorithm
 - RAPL supports $1/8W$ steps
 - Move measurements in initial phase of compute kernel
 - Use better fitting search algorithm for target settings
- Include detailed knowledge about individual CPUs
- Move from proof of concept to more refined solution.
- Integrating with OS / System software / Software-Stack



- Brief Overview of the GREMLIN Framework
 - Emulation helps to understand Systems
 - Possibility to evaluate developments for Future Systems
 - Approach to test System- & Application-Software
- Power studies for power limited system
 - Rebalancing Power is feasible at different scales.
 - Proof of concept shows how to save Power, Energy & Time.

<http://www.nm.ifi.lmu.de/~maiterth/>

LMU

LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

Matthias Maiterth

<http://www.nm.ifi.lmu.de/~maiterth/>

