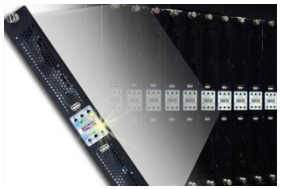
Enhanced Power Measurement with MEGWARE SlideSX®

Thomas Blum – Senior HPC Engineer

ENA Energy-Aware High Performance Computing HPC

HPC developments

- Compute Node Chassis since 2001: Slash2®/Slash5®/Slash8®
- ClustSafe® PDU since 2003: CS12/CS18
- Rack Display for Monitoring: RackView®
- Software for Cluster Management/Monitoring: ClustWare®
- The Green IT Solution: Direct Liquid Cooling with ColdCon®



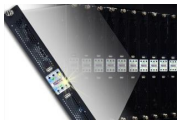


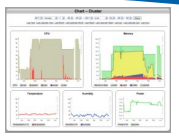

Engineering for HPC solutions

Server grade IT components + Partner R&D

Mechanical – Electrical Engineering
Firm-/Software Development

MEGWARE HPC Products


MEGWARE research activities

Current projects overview

- FaST Project: BMBF funded project on „dynamic topologies in highly scalable environments“
- DIR project: Dynamisch integrierbare Rechenbeschleuniger
- ETP4HPC SME member: European technology platform for HPC
- PRACE 3IP PCP: Whole System Design for Energy Efficient HPC

  MEGWARE selected for Phase I – Design Phase

ClustSafe®

Power Distribution Unit

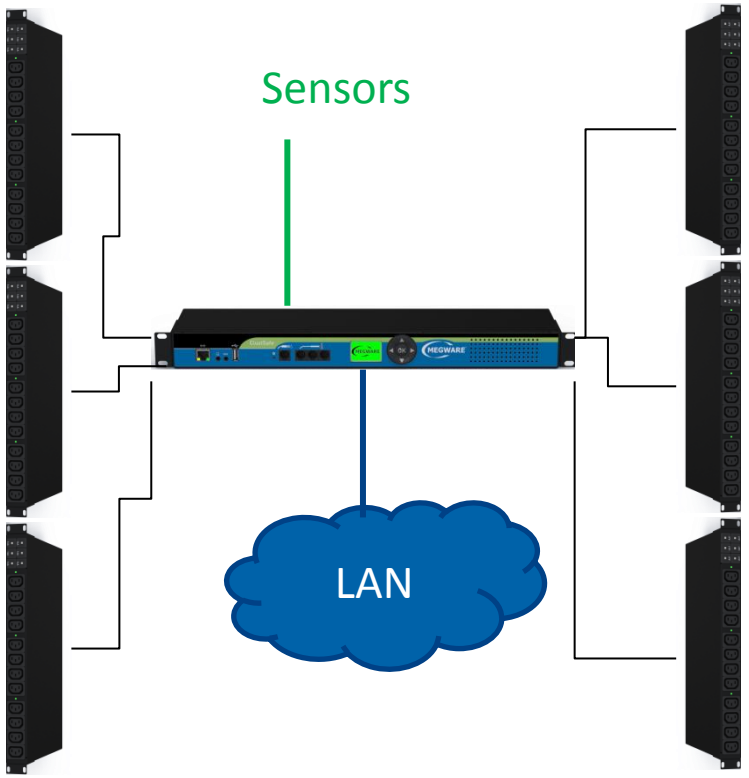
- new version comes with serveral port outlet units and a single rack control unit
- Outlet units for vertical integration in the rear area of the racks
- Rack control unit takes only 1RU per rack and controls up to 8 ClustSafe outlet units with 96 ports
- RCU with ARM based µC and own board design
- Linux based firmware for easy feature and maintenance upgrades supporting web interface via https, snmp and CLI
- preserve functionality from previous versions: filters, power measurement per port/phase, sensors, control panel/display

ClustSafe®

Power Distribution Unit

Rack Control Unit

CS Outlet Units with 12 ports
up to 3 x 32A inlet power
and power measurement



SlideSX®

HPC Compute Platform

- Based on standard server components and MEGWARE mechanical design and electrical engineering
- Flexibility in compute architecture and cooling (air/liquid direct)
- Optimized for maximum air flow and increased inlet temperatures (no front IO backplanes)
- Efficiency optimized for HPC workloads and reliability through redundancy
- Unique features: per node DC power measurement, feature extensions through firmware upgrades

SlideSX®

HPC Compute Platform

Standalone HPC Servers Slash 5/8

19“ density:
10 nodes @ 7U

19“ density:
20 nodes @ 9U

Integrated HPC Compute Platform SlideSX®

Intel® Cluster Ready

19“ density:
10 nodes @ 5U

SlideSX®

Design and Dimensions

Front View - control panels and highly perforated bezel

5U height
up to
8 chassis
per
42U rack

800mm
depth

19“ wide

SlideSX®

Design and Dimensions

Rear View - up to 5 power supplies and management port

redundant load balancing power supplies with 1620W or 2000W

remote manage-ment and monitoring

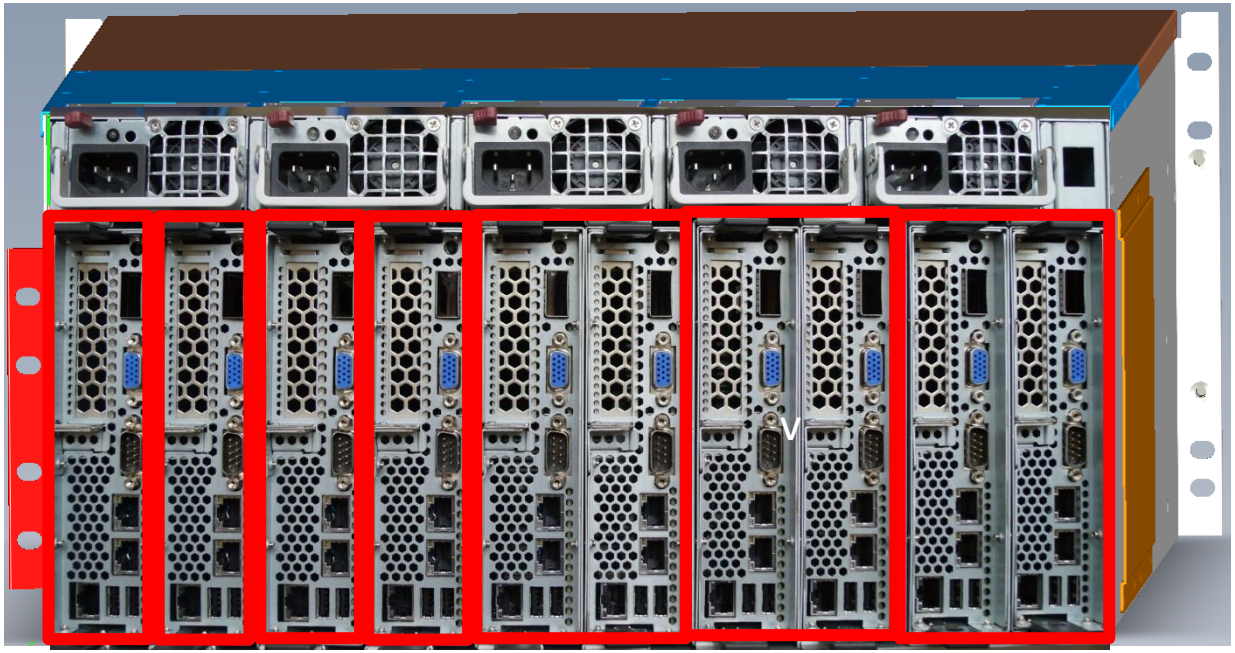
flexible configuration – 5 double or 10 standard compute nodes in 5 segments

SlideSX®

Design and Dimensions

Rear View - up to 5 power supplies and management port

redundant load balancing power supplies with 1620W or 2000W



remote manage-ment and monitoring

segments allow mixed configuration of CPU compute nodes and accelerated double compute nodes

SlideSX®

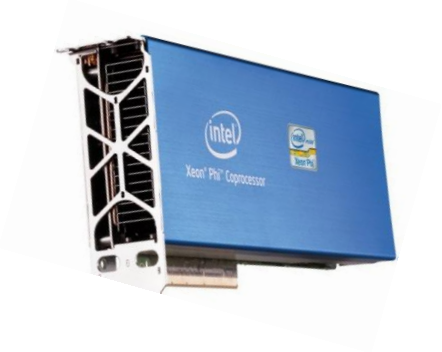
Compute Node options

- CPU compute node specs:
  - Dual Intel Xeon CPU Socket R – SNB/IVB up to 12 Cores
  - 8 DIMM slots for up to 256GB DDR3-1866 RAM
  - Dual Gigabit Ethernet onboard, dedicated IPMI
  - up to 2 x 2.5“ hard disks– mechanical or SSD
  - x16 PCI-E Gen3 full height riser slot for add-on PCIe cards
  - High speed networking options:
    - onboard Mellanox Connect-X3 QDR/FDR PCIe Gen3
    - Mellanox Connect-IB FDR
    - Intel True Scale QDR IB
    - Extoll 2D/3D
    - 10GE/40GE

SlideSX®

Compute Node options

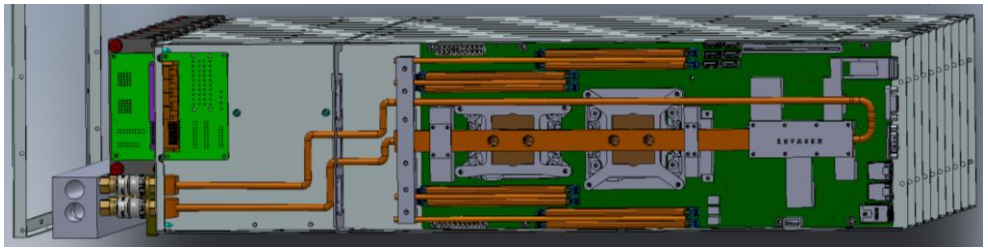
- Accelerated double compute node specs (available in Q3/2014):
  - Dual CPU Intel Xeon E5-2600v3 Socket R3
  - up to 512GB DDR4-2133RAM
  - Gigabit Ethernet onboard, dedicated IPMI
  - 2 x x16 PCI-E Gen3 full height riser slots for add-on PCIe cards
  - High speed networking options:
    - onboard Mellanox FDR IB
    - Next Gen Mellanox IB (when available)
    - Intel True Scale QDR IB
    - Extoll 2D/3D
    - 10GE/40GE
  - up to 2 NVIDIA Tesla or Intel Xeon Phi (passive cooled, PCIe FF)

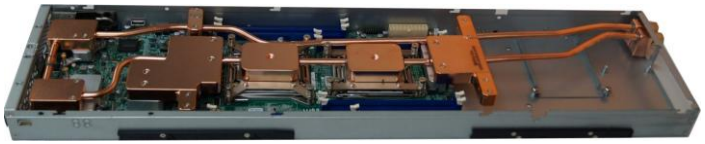


SlideSX®

Cooling options

- standard air cooling for power supplies and compute nodes with fans inside compute node drawer
- direct liquid cooling for inlet temperatures up to 60°C enables reuse of waste heat
- same chassis for both options, air cooled systems can be enhanced for liquid cooling
- ColdCon® with all onboard components direct liquid cooled including CPUs, VR, DIMMs and chipsets including IB connector (Q3 2014)





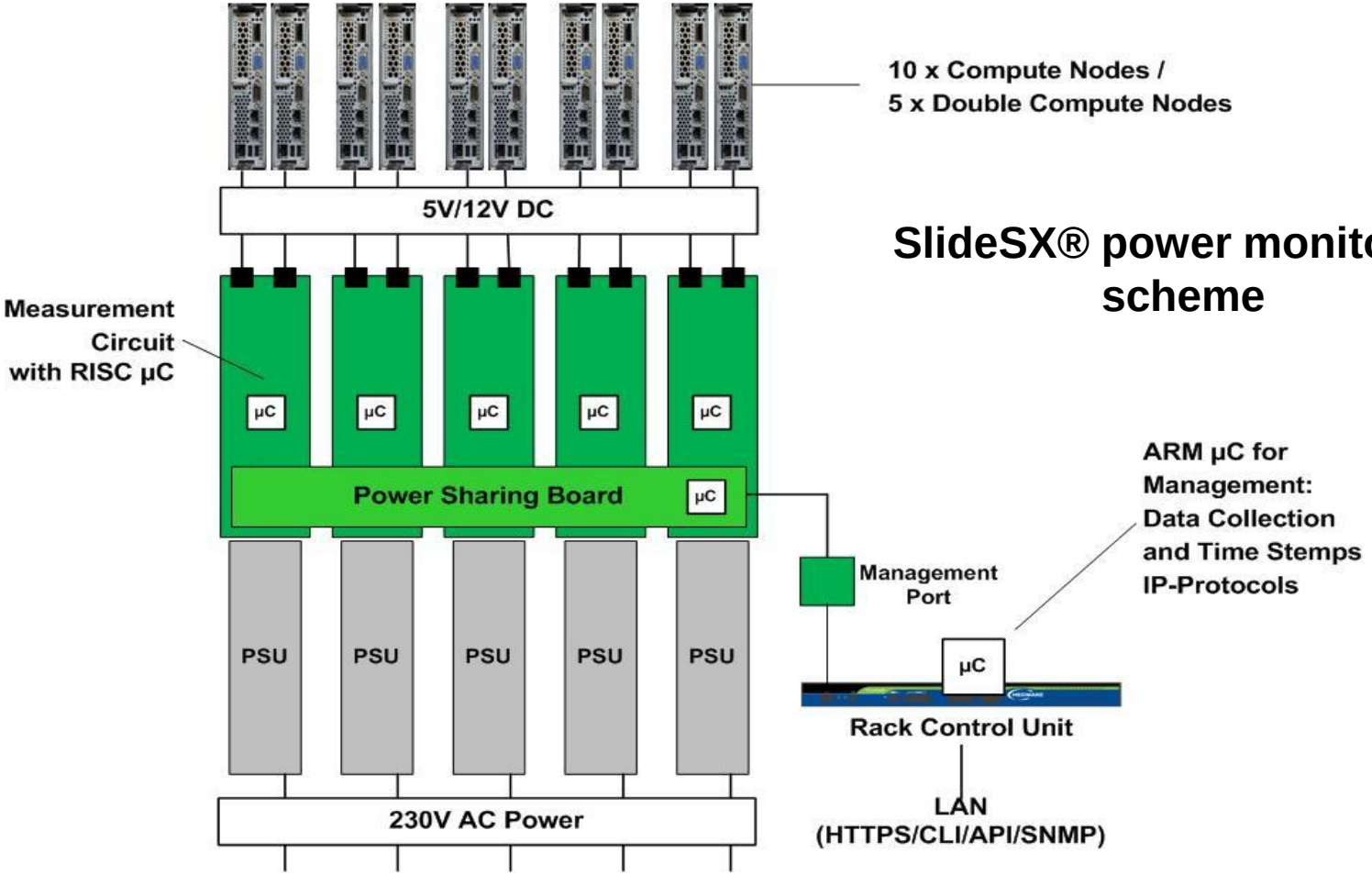
SlideSX®

Power sharing and measurement

- up to 5 power modules with 1620/2000W in N+1 redundancy
- 80+ Platinum certified with 94% high efficiency
- standard configuration: 3 + 1 for 10 compute nodes
- performance configuration: 4 + 1 for 5 double compute nodes
- PMBus Management and additional control functions engineered by MEGWARE
- Out-of-Band Monitoring via RCU
- detailed and accurate per node power monitoring with sampling rate up 5000/sec on DC side for 5V and 12V

SlideSX®

Power sharing and measurement



SlideSX® power monitoring scheme

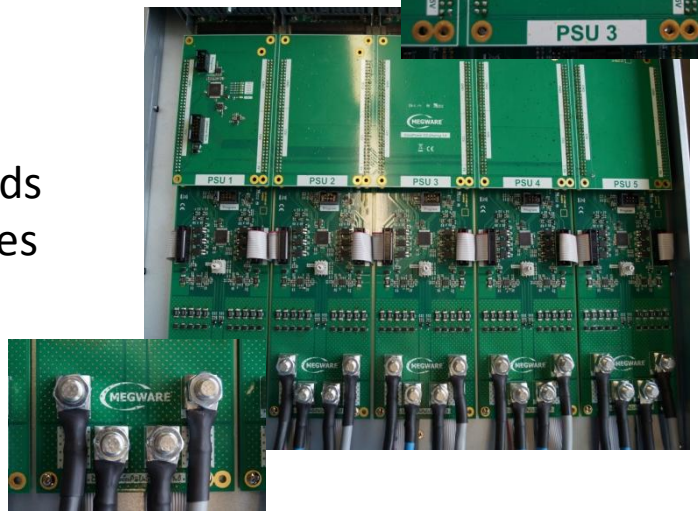
SlideSX®

Power sharing and measurement

PSU module

Power Sharing Board

Node Power Boards supporting 2 nodes incl. power measurement



SlideSX® Backplane

SlideSX®

Chassis Management

8 x SlideSX Chassis with 80 servers

RCU – Rack Control Unit

LAN

one RCU per rack for management and monitoring of the servers

SlideSX®

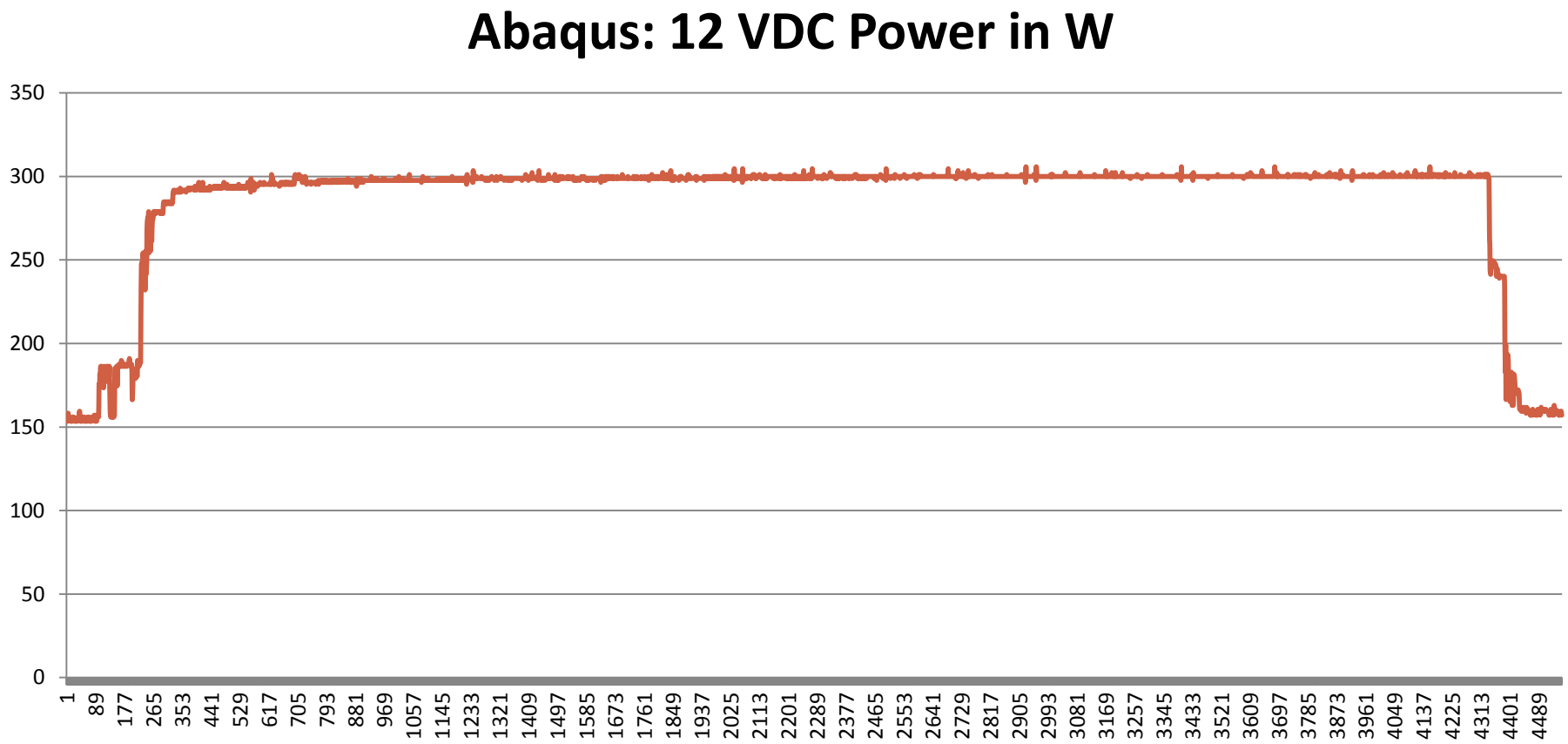
Power sharing and measurement

Test-Bed configuration:

- MEGWARE SlideSX® compute platform + RCU
  - SlideSX® 5/12 VDC power measurement facility
    - 15000 samples/s split into 15 x 1k measurement points = 15 records per second
  - 1 x SlideSX® compute node
    - 2 x Intel Xeon E5-2670v2 10C 2.5Ghz 115W TDP
    - 8 x 8 GB DDR3-1866 DIMM
    - 240GB Intel SSD
    - Turbo Mode Enabled / HT Enabled (HT cores unused)

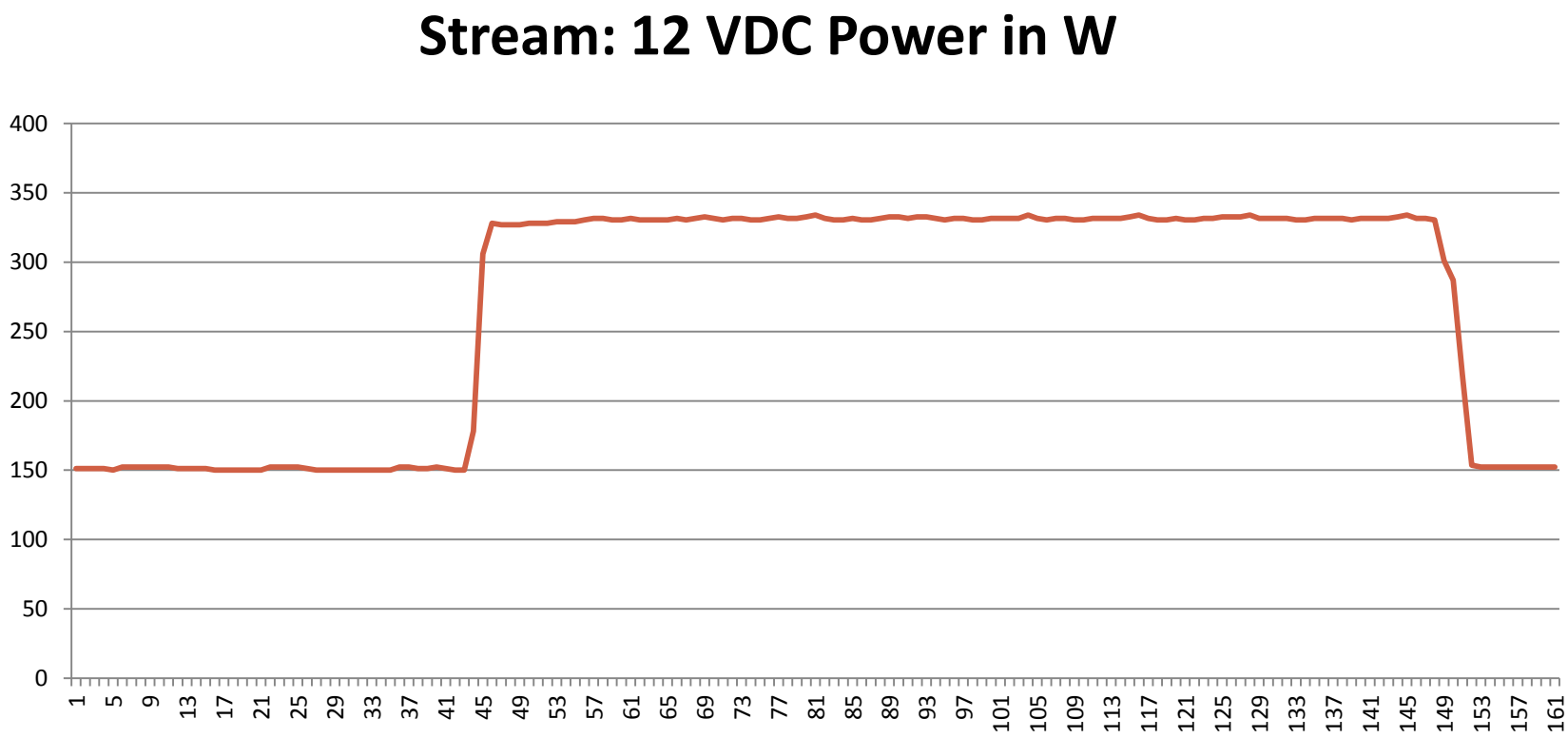
SlideSX®

Examples: DC power measurement on a single node



SlideSX®

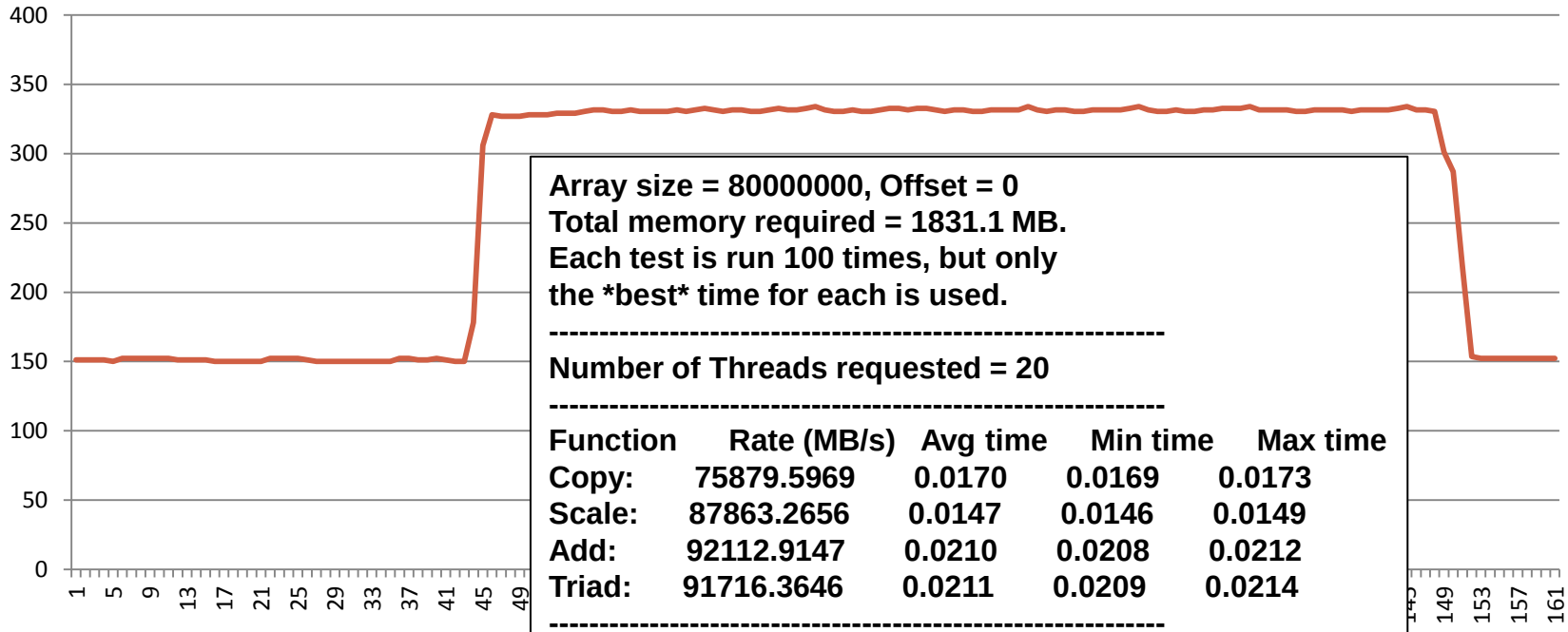
Examples: DC power measurement on a single node



SlideSX®

Examples: DC power measurement on a single node

Stream: 12 VDC Power in W



Array size = 80000000, Offset = 0
Total memory required = 1831.1 MB.
Each test is run 100 times, but only
the *best* time for each is used.

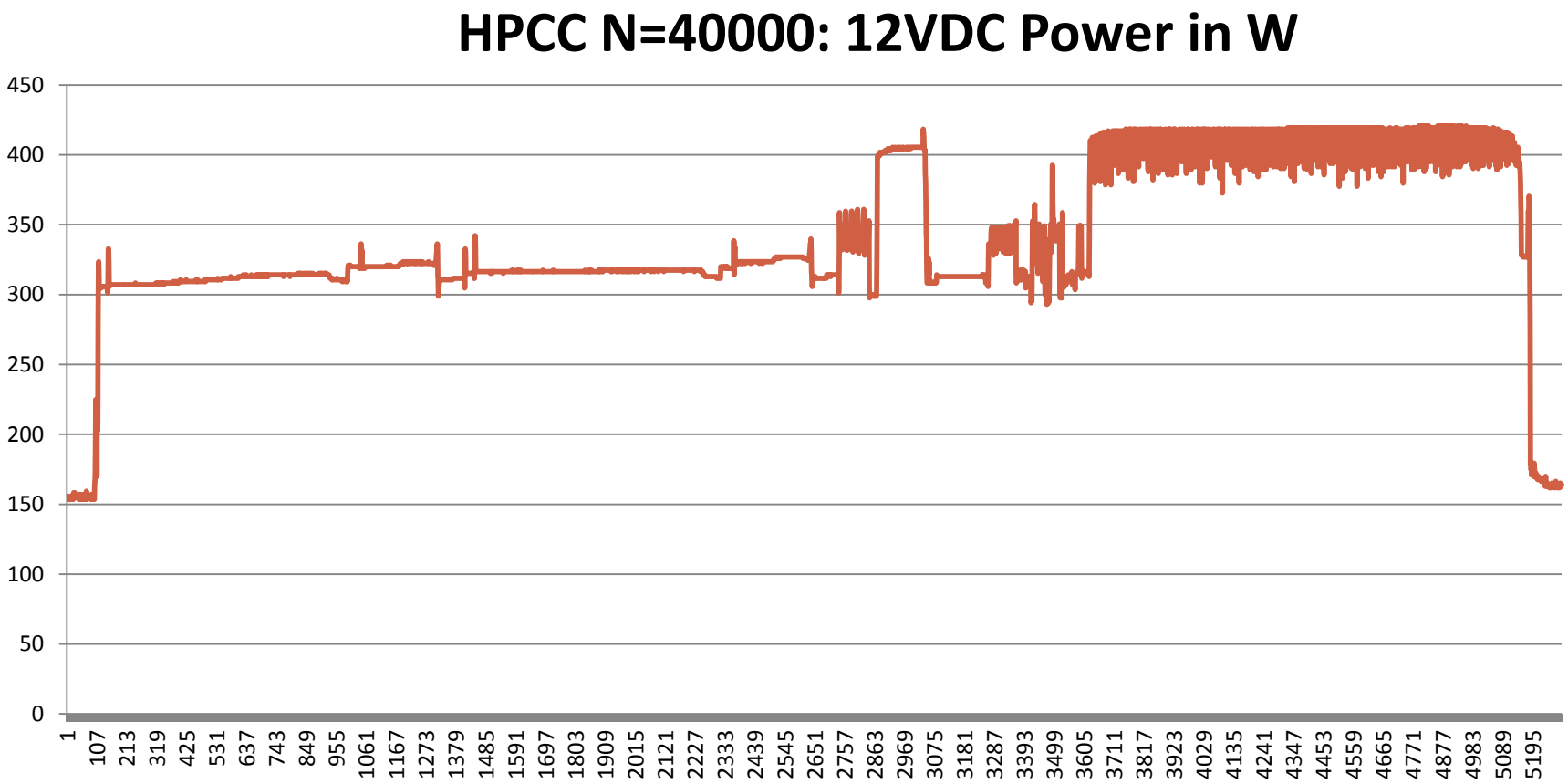
Number of Threads requested = 20

Function	Rate (MB/s)	Avg time	Min time	Max time
Copy:	75879.5969	0.0170	0.0169	0.0173
Scale:	87863.2656	0.0147	0.0146	0.0149
Add:	92112.9147	0.0210	0.0208	0.0212
Triad:	91716.3646	0.0211	0.0209	0.0214

Solution Validates

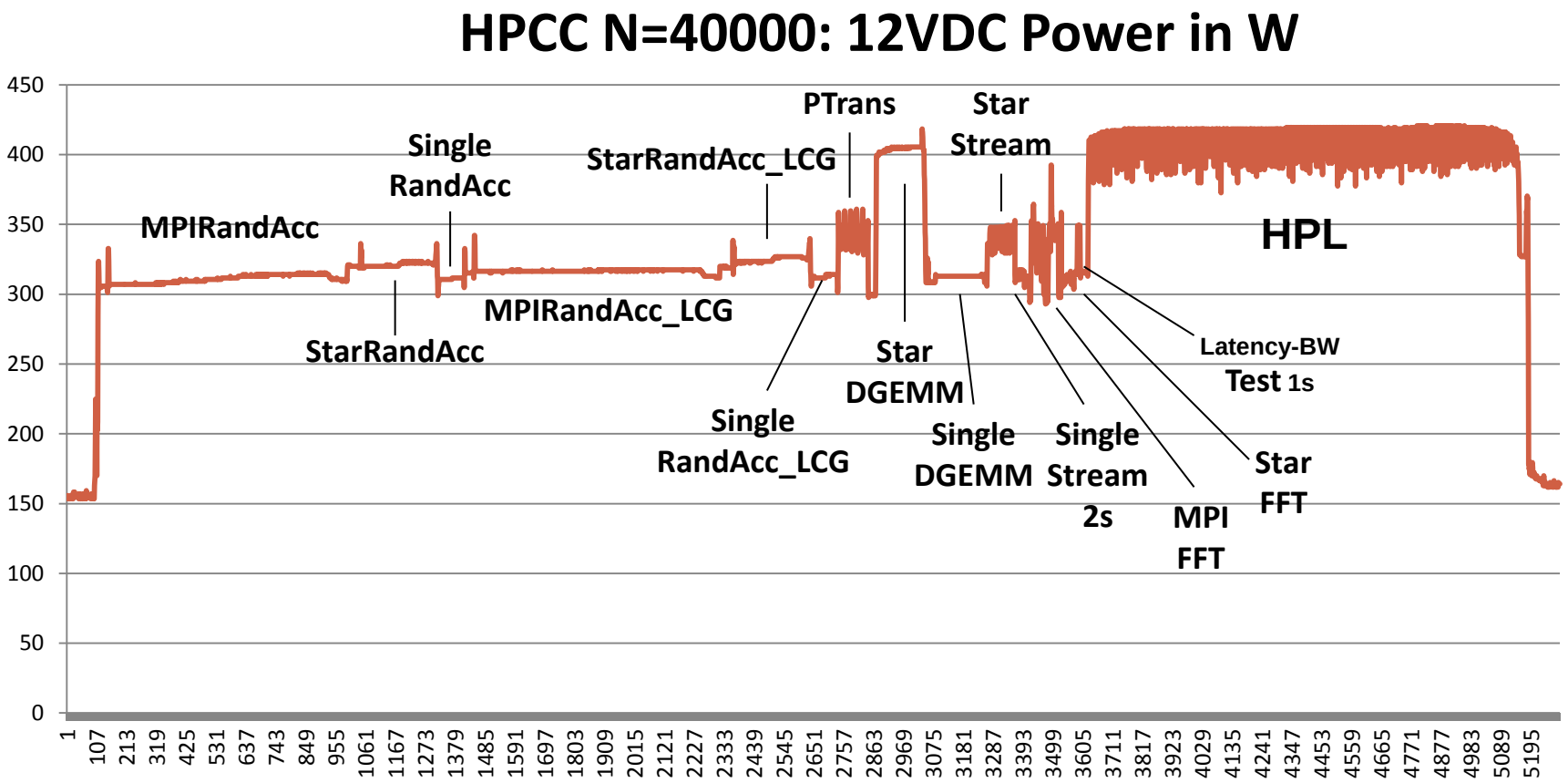
SlideSX®

Examples: DC power measurement on a single node



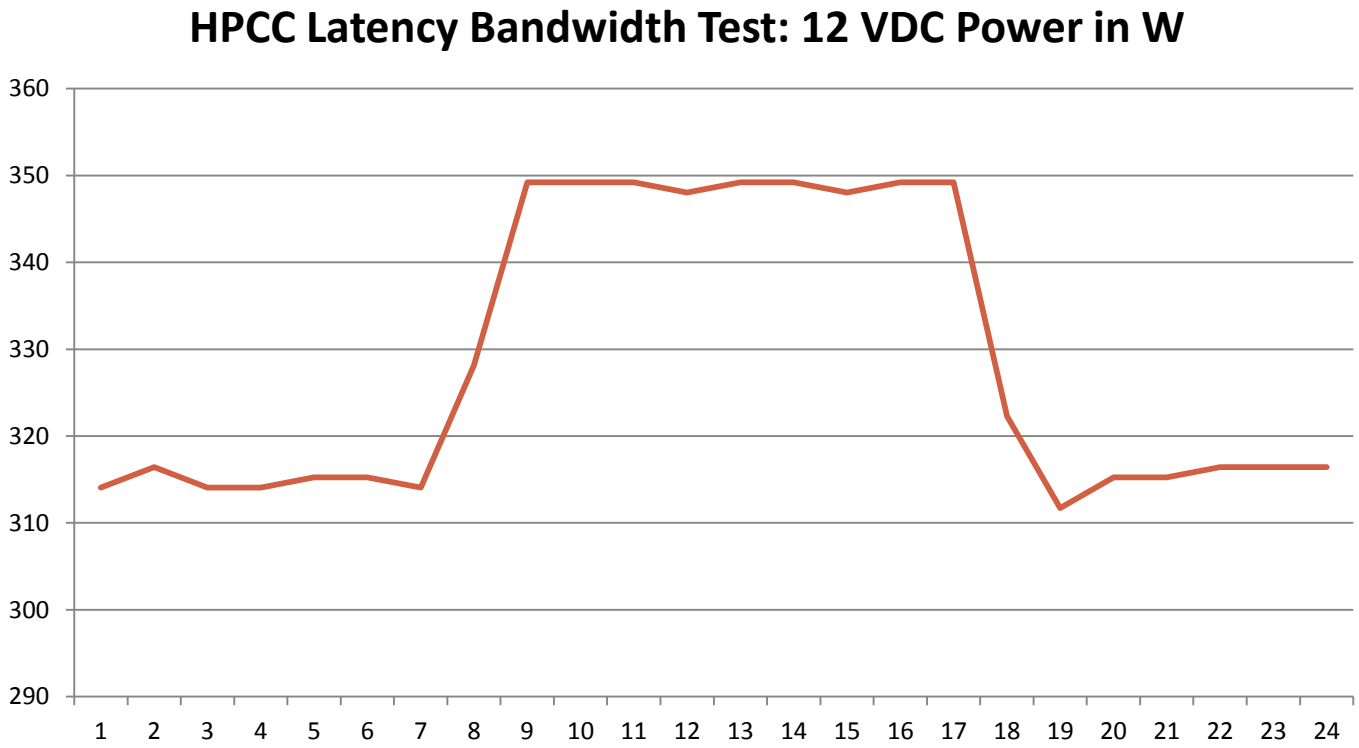
SlideSX®

Examples: DC power measurement on a single node



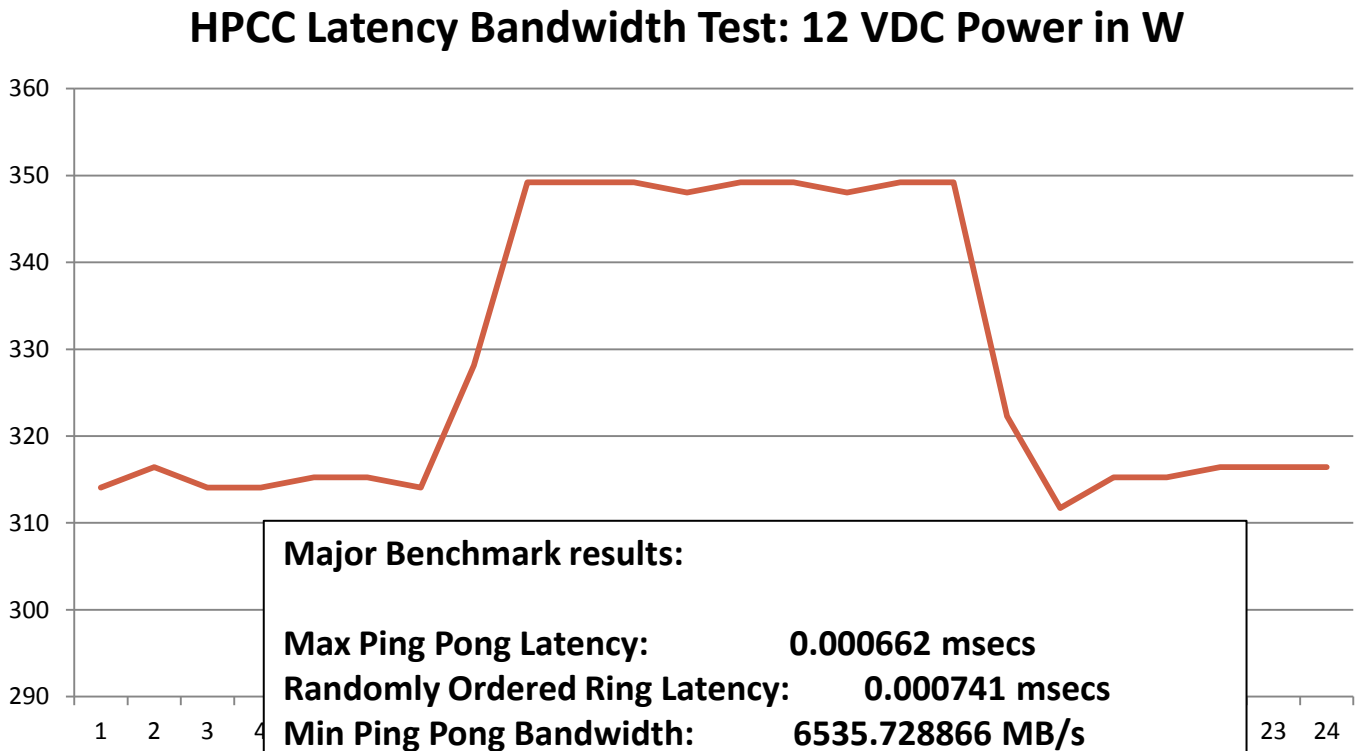
SlideSX®

Examples: DC power measurement on a single node

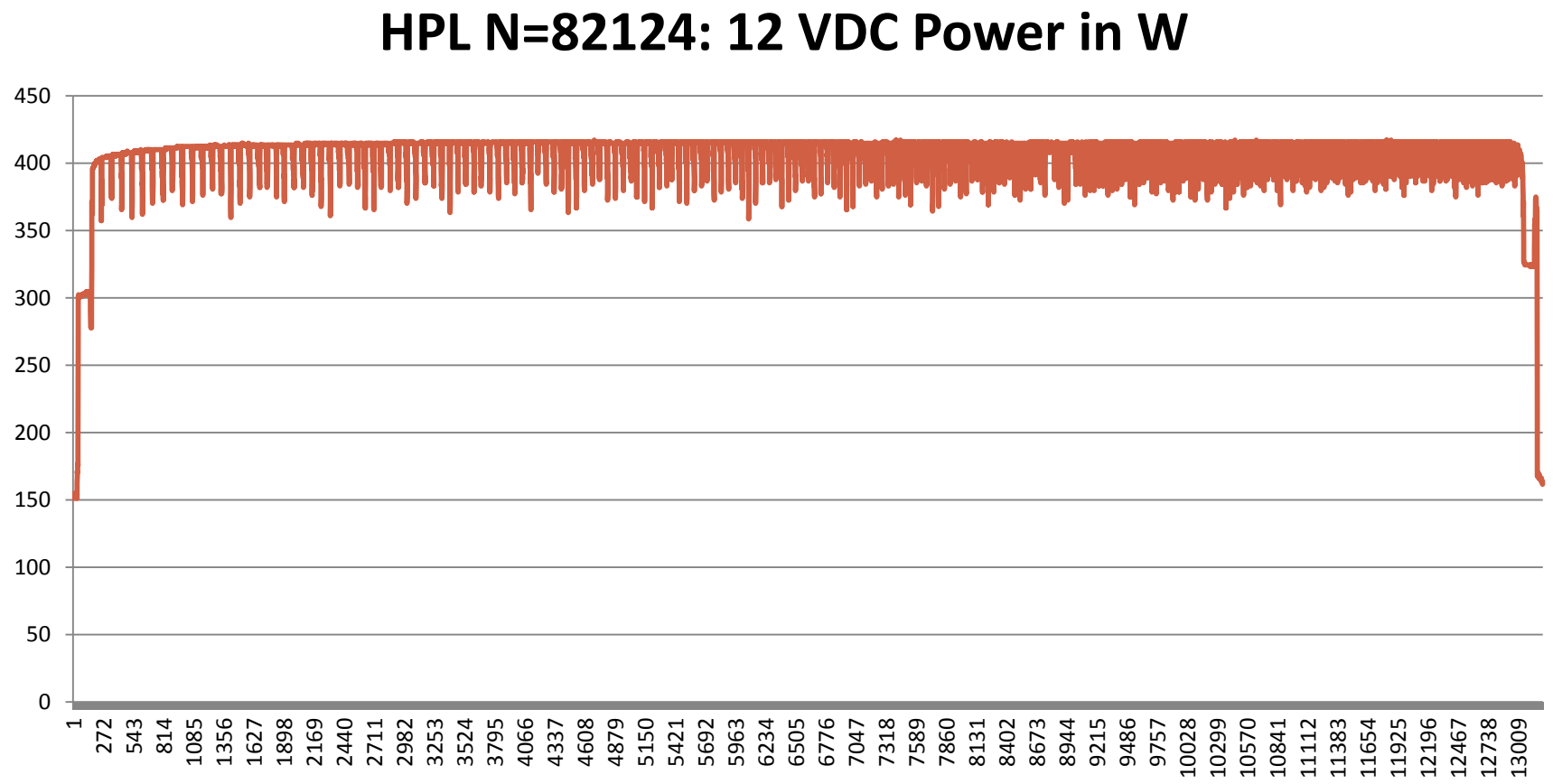


SlideSX®

Examples: DC power measurement on a single node

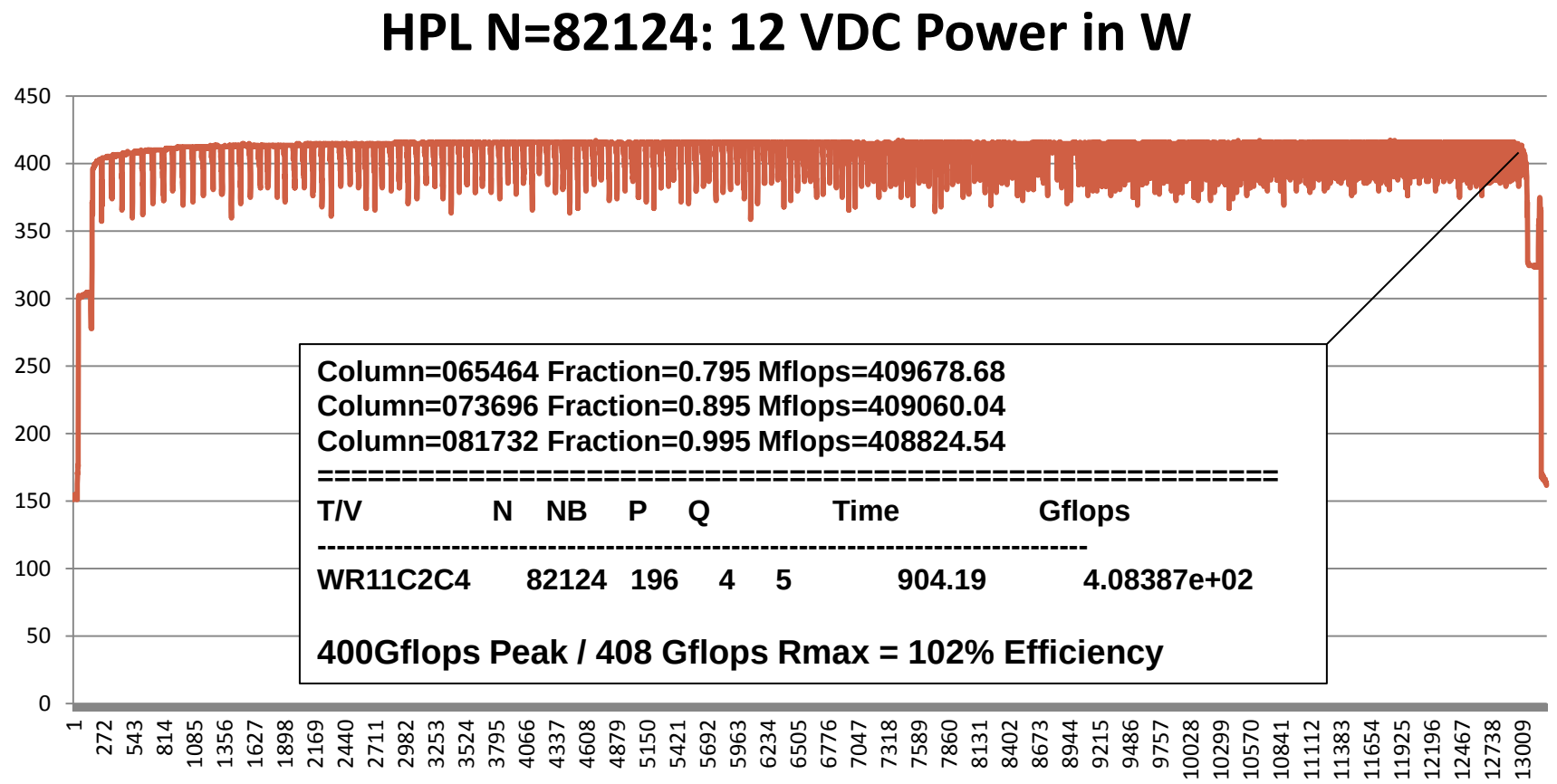


Examples: DC power measurement on a single node



SlideSX®

Examples: DC power measurement on a single node



SlideSX®

Conclusion and Outlook

- DC Power monitoring will be available with fixed sample rate of 5K samples for 1 record per second for all compute nodes per SlideSX® chassis
- for research projects more fine grained measurements can be achieved
- Firmware can be changed or modified to specific needs
- Definition of Power caps per compute node for future firmware release

Thank you for your attention!

Q & A

thomas.blum@megware.com