



Experiences Developing and Deploying Per Node Power Monitoring at Scale

HPC Power Management 2015
Annapolis, MD

Penguin Computing: who are we?



Founded in 1999.

Focused on Linux solutions across all disciplines.

- Installed base: 10K systems in over 40 countries.
- Customers include: AOL, NASA, Caterpillar, Lockheed Martin, Boeing, Life Technologies and the US Navy.

Provider of HPC solutions, including:

- Turn-key clusters with compute, storage and interconnect
- Scyld ClusterWare – provisioning and management software
- POD (Penguin On Demand) – HPC as a Service

Discussions with Jim Laros who wanted:

- “Detailed” measurements of different parts of a system
- No impact to host CPU, memory, etc.
- Fast sampling (many samples per second)
- Network connected for out-of-band access
- Cheap enough to instrument a complete cluster
- Small enough to allow nodes in rack deployments
- Support “Power API” development and implementation

Idea that ATX Power interfaces provide:

- Discrete power rails that can be measured
- Each rail can be mapped to specific components

Development of the Idea



ACS713 integrated Hall effect current sensor

- 0-20A and 30A Ranges
- Output is ratiometric to VCC

Create a PCB with power connectors and in-line measurement

Use a 32-bit embedded microcontroller

3.5in hard drive size (fit in available drive bay)

Microcontroller was difficult to design with

- Need for an OS or supervisor
- Component selection and layout for Ethernet port
- Network stack for microcontroller
- Switch to BeagleBone and Linux

Multi-channel board was inflexible

- Multiple PCB's would have to be designed and sourced
- Low volumes (10-100 vs 1000)
- Create a single rail sensor and different wire harnesses

Delivery of the idea (v1.0)



Added as instrumentation on existing cluster

- 104 node AMD APU cluster at Sandia
- Seven rails measured:
 - ATX 12V, 5V, 3.3V, 5Vsb
 - CPU 12V
 - HDD 12V, 5V
- CPU and RAM on different rails
- Simple C program to collect measurements
- 1000 samples per second
- Developed diskless NFS root config for BeagleBone
- Custom kernel required to access A2D devices

Rail based measurements allowed creation of PCI-Express riser to measure add-in cards

- 12V and 3.3V rails on PCI-Express connector
- 7/8in high riser and standoffs compatible with 4U chassis
- Typically used to measure total GPU power

Planning for v2.0

- Upgrade from 10-bit to 16-bit ADC
- Add temperature measurement Type-K thermocouples
- Expandable up to 60 rails
- New sensors for low current rails (<5 Amps)

Delivery of v2.1



Multiple customer deployments

- PNNL, Sandia, ORNL, VA Tech, RENC

Kernel now uses Device Tree

Software uses Lua for control and configuration

- Support for multiple sensors
- Latest Power API interface

Haswell/Broadwell systems now shipping

- Independant CPU and RAM rails
- 2U chassis
- 4U chassis supports 3 GPU. Uses 2x PI systems

New sensors using 10A Hall effect device

- PCI-Express in particular benefits from this

Fan control expansion

- Measurement and control of fan PWM and Tach

V3.0 will probably be software in BMC

- Power sensors now appearing in IPMI SDR
- Open source BMC firmware available
- Leverage intelligent VRM controllers on RAM and CPU rails

Power Insight Watt Metering



Flexible power monitoring platform

- Sensor modules to instrument power rails

- Main board with analog inputs

- Expansion boards with K-type Thermocouple inputs

- Future expansion boards (analog or other)

- Embedded Linux controller (BeagleBone Black)

PCI-Express riser

- Sensors for power rails from PCI-Express slot

- 12V and 3.3V

Power Insight Watt Metering (cont)



Direct measurement of actual voltage and current

- Hall-effect current sensors or shunt resistors

- High sample rates (>1000 S/sec)

- 15 channels on main board

- Additional channels on expansion boards

Completely non-invasive

- BeagleBone Black does all processing

- Local 10/100 Ethernet port

- USB and Serial connections to host for coordination

Power Insight Watt Metering (cont)



Small enough to fit inside standard chassis

- Carriers are 3.5in hard drive sized

- Sensors in-line in power harnesses

- “Close the lid” and in-situ testing

Affordable enough to instrument all nodes

- Already deployed in multiple customer clusters

- 104 node prototype

- Four cluster from 16 to 36 nodes

Thank you for listening.

We're happy to answer any questions.