OCP SUMMIT

March 20-21
2018
San Jose, CA

OPEN. FOR BUSINESS.

UPDATING THE OCP COMPUTE VOLTAGE STEP RESPONSE SPECIFICATION

John Nguyen/Principal Product Manager/
Penguin Computing

Agenda

PENGUIN COMPUTING

OPEN PLATINUM

OPEN Compute Project™ SOLUTION PROVIDER

- About Penguin Computing (& why we care about OCP standards)
- Voltage Instability Under Certain Workloads
- Cluster Configuration
- RCA
- Cap Shelf
- Future Upgrade
- Discussion

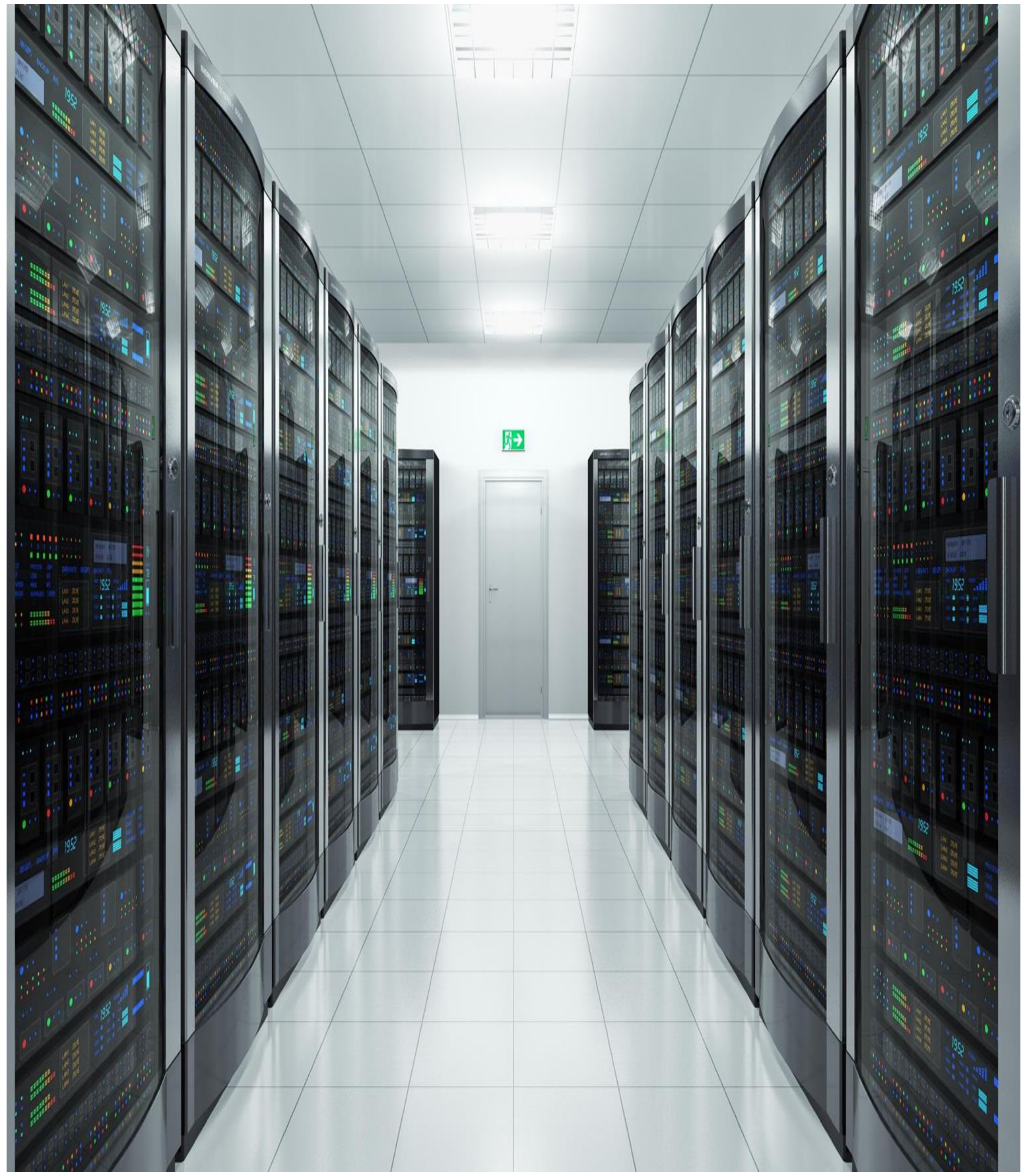
About Penguin Computing

- U.S.-based 20 year old, global provider of HPC hardware, software, and services
- Home to Scyld® Beowulf cluster software & bare metal HPC on cloud Penguin Computing On-Demand™
- Over 300 OCP racks delivered to date based on Tundra™ Extreme Scale design*
- Platinum OCP member, Penguin CTO Phil Pokorny is HPC representative of the OCP Incubation Committee

* OCP Inspired, for discussion Q2

Voltage Instability Under Dynamic Workloads

- Power shelf can't respond quick enough during the rapid transition of power state.
- 64 nodes all transition from medium to full power in same microsecond, 12V rail sagged
- 64 nodes all transition from FULL power to IDLE in same microsecond, 12V rail surged
- High Speed, Low Latency Fabric
  - Work load synchronized within microseconds between nodes
- Initial application was LAMPS, reproducible in other workloads



Cluster Configuration with Voltage Instability

- Relion OCP1930e*
  - Dual Intel Xeon E5-2695 v4
  - 128GB DDR4-2400MHz (8x16GB)
  - Intel Omni-Path 100
  - Diskless Boot
  - Air-cooled^
- 64 Nodes per Rack

* Spec in submission to OCP Foundation for technical review - Q2’18

^ image depicts liquid-cooled sled; air-cooled sled used in case study, and also available

Cluster Configuration with Voltage Instability

- Open Bridge Rack (40 OU), Single Zone
- 26kW, N+1, IEC60309 560P7-3P, 5Wire, 277/480V, 60A
- Vertiv NetSure Rectifier and Power Shelf

Cluster Configuration with Voltage Instability

OPEN Compute Project™ SOLUTION PROVIDER

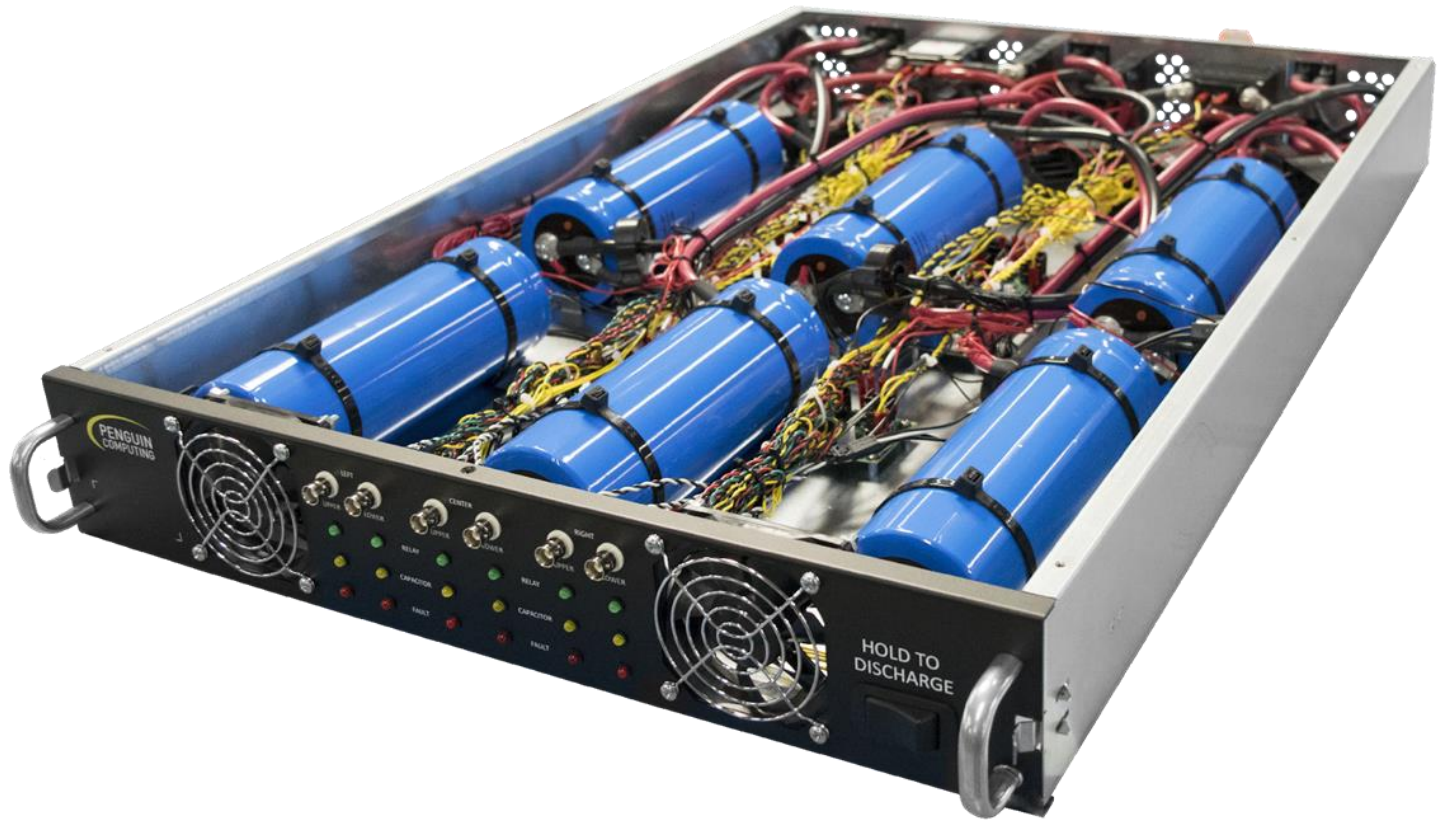
- At high node density and correlated workloads, can exceed slew rate spec for power shelf
- Symptom shows up at scale, seen at 576 nodes; not seen at smaller node counts
- It's an HPC thing - 50% increase in nodes

Root Cause Analysis

- Intel datasheet says CPU can draw more power than stated. Exceed TDP for up to 4 milliseconds
- Cluster power specification based on calculated power budget off steady TDP value -- We knew it can exceed, but by how much?
- We were put into a position to consider over-provisioning rack power
- Over-Engineering and investigating which specific RACK configuration will create a problem, would be wrong decision to make all things considered
- But….**it happened!**

Capacitor Shelf

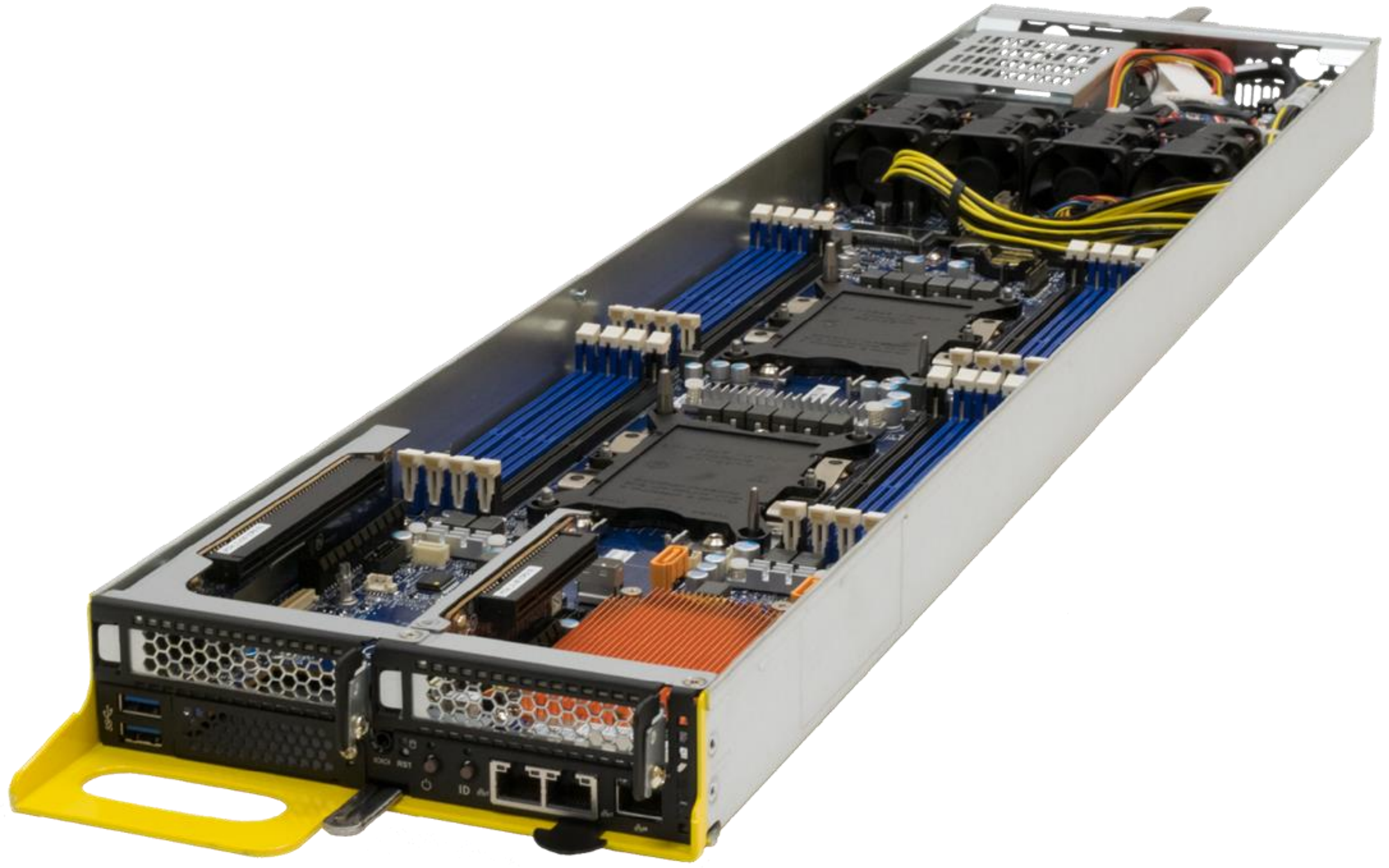
- Slew rate issue, not total power
- Prototype w/ capacitor directly to bus bars, proved it can be effective
- Designed additional safety feature, charge/discharge circuit
- 1 Farad (1 million uF) shelf
- It works, and passed FCC/UL.
- New Design: Capacitor Shelf “Power Buffering Solution”



Future Upgrade

OPEN Compute Project™
SOLUTION PROVIDER

- Incorporated the capacitor shelf into each individual node so rather than one large capacitor, it's many smaller capacitor per node.
- Discussion with power supply MFR, to discuss increase slew rate for HPC applications for their technology
- Recommend update to OCP Open Rack power specification to increase slew rate spec. (~2000A /µsec)



Discussion

OCP
SUMMIT