The bridge to possible

Reducing Energy Consumption with Intel Xeon Power Management Technologies

Danny Cartagena
Principal Engineer
Intel - Data Center and AI Group

BRKCOM-1329

#CiscoLive

Cisco Webex App

Questions?

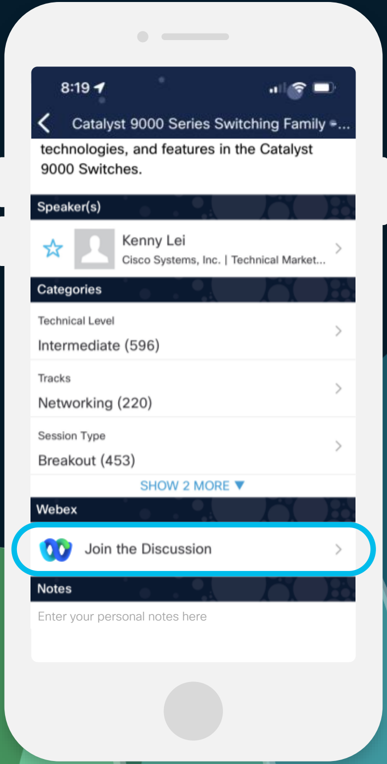
Use Cisco Webex App to chat
with the speaker after the session

How

1. Find this session in the Cisco Live Mobile App
2. Click “Join the Discussion”
3. Install the Webex App or go directly to the Webex space
4. Enter messages/questions in the Webex space

Webex spaces will be moderated
by the speaker until June 7, 2024.

https://ciscolive.ciscoevents.com/
ciscolivebot/#BRKCOM-1329


8:19

Catalyst 9000 Series Switching Family ...

technologies, and features in the Catalyst 9000 Switches.

Speaker(s)

Kenny Lei
Cisco Systems, Inc. | Technical Market...

Categories

Technical Level
Intermediate (596)

Tracks
Networking (220)

Session Type
Breakout (453)

SHOW 2 MORE ▼

Webex

Join the Discussion

Notes

Enter your personal notes here

Data Center Sustainability Focus

SOC & Platform Energy Efficiency

- Microarchitectural Power Optimizations
- Load Line Linearity
- Idle Power Improvements

Operating Carbon Footprint & Measurement

- Power Telemetry
- Intel Platform Monitoring Technology

Rack & DC Optimization

- Data Center Optimization
- Liquid Cooling Scaling
- Immersion Cooling Warranty

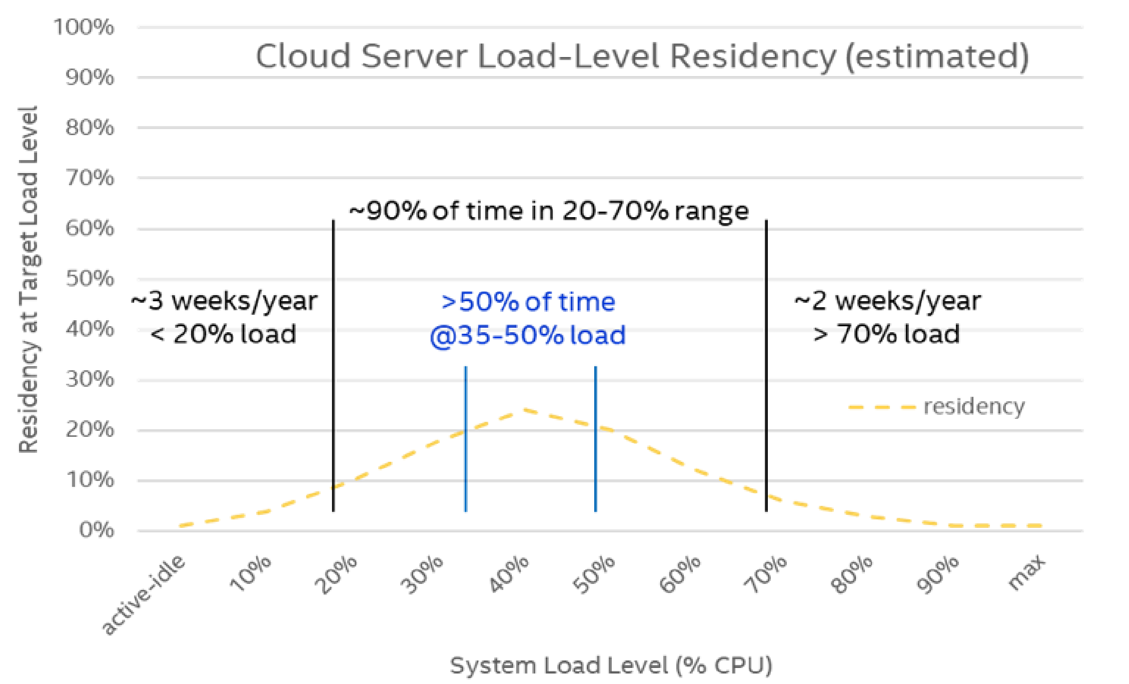
Life Cycle Optimization

- Extended Servicing
- Circularity
- Modular Hardware System (DC-MHS)

Energy Efficiency & Operational Carbon

Embodied Carbon

Cloud Example of Varying Server Utilization



Most servers operate between 35% to 50% utilization

Power management features improve energy efficiency

Selectively Deploying & Configuring Power Management

Higher Utilization / SLA

Server Activity / Use Case	Example Power Management Features
Over Subscribed: Power Capping	Psys, Socket & DRAM RAPL, Pmax, etc.
High: Max Perf/Utilization	UFS (uncore/core freq sharing), Turbo, Energy Effiicient Turbo
Typical	UFS, P-states, HWPM, Core C-State (CC1), PCIE L1, etc.
Low Utilization: Performance Idle	Core C-State (FC1E), Active Idle Mode, Optimized Power Mode, UFS limits
Low Utilization: Active Idle	Core C-State (CC6) Active Idle Feature, Optimized Power Mode, UFS limits
Low Utilization: Unprovisioned	Dynamic PC6 (DRAM Self Refresh)
Exceess Capacity: Unprovisioned	Dynamic PC6+: PC6 (DRAM Self Refesh) + PC2 optimizations
Installed Not Deployed	S5

Decreased Power Increased Latency

Power Management features have evolved to minimize performance impact, can be enable dynamically (w/o reboot), and have increased configurability

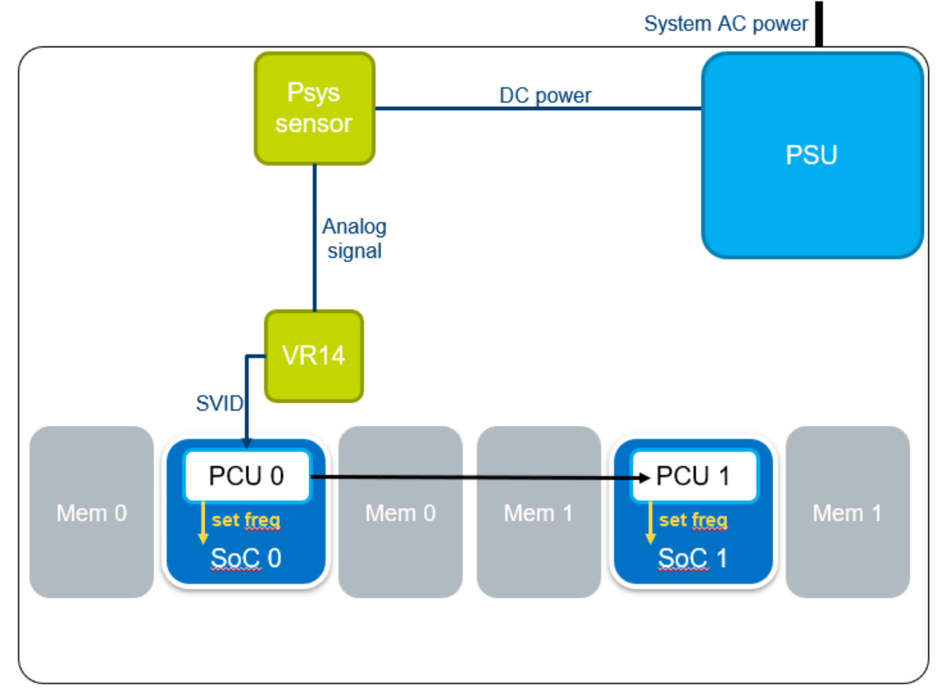
Improved Power Capping via Psys

Rack overprovisioning is key to maximizing compute density

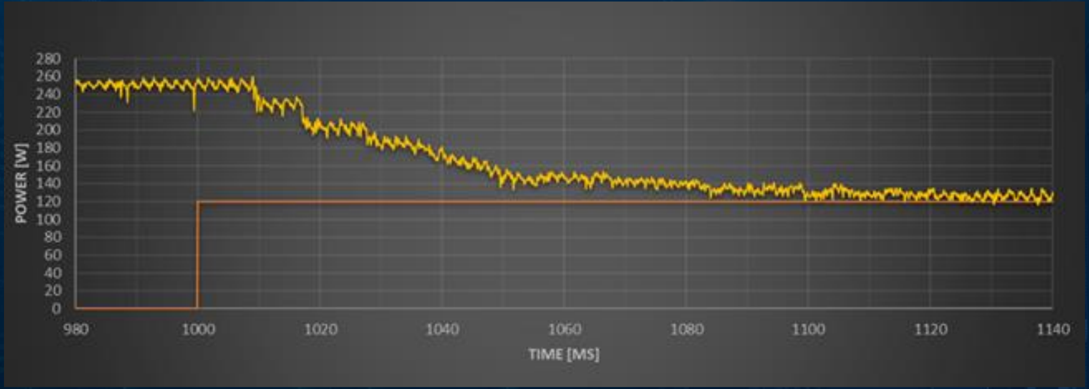
Fast and accurate power capping is necessary to ensure power does not exceed rack limits

Legacy power capping involved sensing of system power via board-level telemetry.   BMC and/or workload orchestration SW then modulated CPU power.

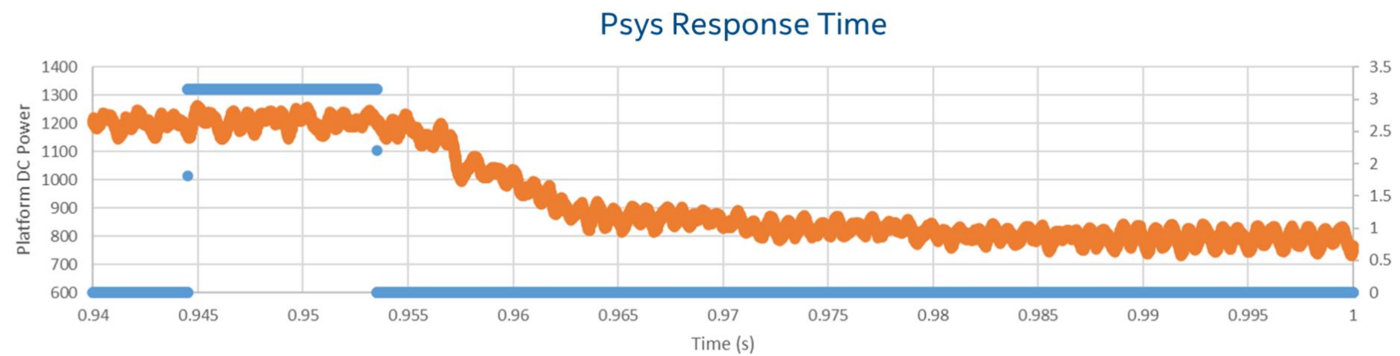
Xeon Psys feature enables CPU-based closed loop control by sensing total system power and actuating CPU frequency to maintain power limits



Example Psys Response Times



Fast Node Manager Power Capping. ~25ms response & ~100ms setting time


Psys Response Time

4th Gen Xeon Psys response as fast as ~5ms with settling time as fast as ~16-25ms

Intel Measured Results on Intel Reference Systems. Results may vary.

SOC & Platform Power Management Features for Energy Efficiency

Nearly 90% of data center carbon emissions come from operating and cooling IT equipment. This collection of power management features optimize the energy consumption of the SOC and Platform to increase the energy efficiency of the data center.

Active Idle Mode

A power recovery strategy which lowers Uncore frequency in low activity scenarios. Customers can set minimal utilization point and thresholds to minimize impact to workload performance. Power savings when customers disable low power states and are in idle condition.

Optimized Power Mode

Intel BIOS option for up to 20% power reduction with minimal performance impact*. Found under CPU-Advanced PM Tuning. New enhancements pending for Birch Stream.

Fast C1e

Per core low power state with better reliability and responsiveness than legacy package level C1e. Significantly lowers exit latency compared to legacy package C1e and does not require package level idleness, enabling voltage reduction in SoC.

New Intel Xeon 6

Core C6 (CC6) Enumeration

Allows the OS to differentiate CC6 and Package C6 (PC6) when using the MWAIT instructions. Enables fine grain control of CC6 and PC6 to select states independently using existing Intel-idle driver interfaces. Lower CC6 exit latency means the OS can increase CC6 requests, improving CC6 residency and resulting in better energy efficiency and TCO.

New Intel Xeon 6

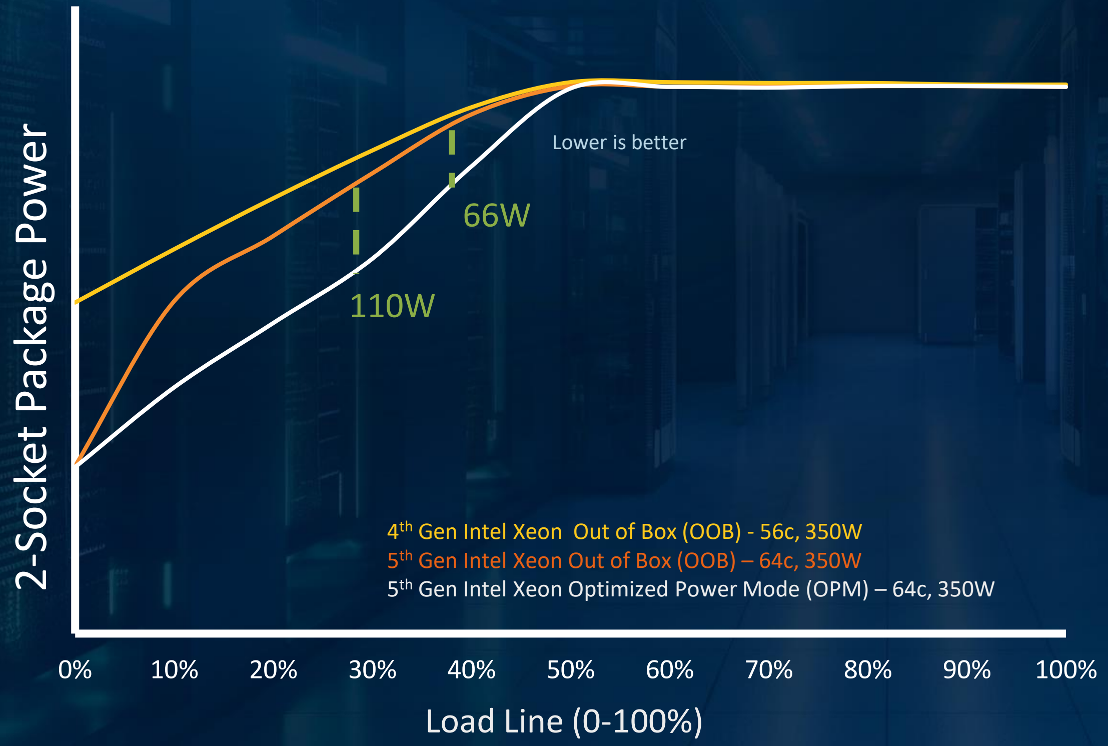
Fast Fabric Frequency Scaling

Provides efficiency improvements by dynamically changing voltage and frequency of compute and IO domains to eliminate down time for transactions contained within a domain

*Impact assessed on SPECcpu, SPECjbb, SPECPower, FIO and NGINX. Higher performance impact possible in workloads more sensitive to lower uncore frequencies at low utilizations

Higher Efficiency at Lower Server Utilization

5th Gen Intel® Xeon® processors lower customer power and cooling costs



5th Gen Xeon Architecture enhancements deliver improved Performance / Watt Out of the Box

Optimized Power Mode
Lower Power & Cooling Costs

P-state Governor = Performance
EPB = Performance
OPM UP = 6, Mesh freq = 1.4G
Mesh max freq = 2.2G

At 30-40% utilization levels, realize up to 110W server power savings

Results measured by Intel on Intel Reference Systems. Results may vary.

Optimized Power Mode 2.0 Now Available

	4th Gen Xeon	5th Gen Xeon
Capability	Optimized Power Mode 1.0	Optimized Power Mode 2.0
What is it?	Easy button for up to 20% power reduction with <5% WL performance impact[*,1]	**Enhanced Optimized Power Mode for up to an additional power reduction with minimal perf impact**.
Components	1. Active Idle Mode (AIM) • Lowering Uncore frequency under low activity scenarios • Minimal Utilization Point (UP) and threshold are chosen as defaults to minimize impact to workload performance • Available OOB with UP=0, leveraged by Optimized Power Mode 1.0 and sets UP=6 2. Fast C1E (FC1E) • New per core low power state with better reliability and responsiveness than legacy package level C1E. • Significantly lowers exit latency compared to legacy package C1E and does not require package level idleness, enabling voltage reduction in SoC • Available OOB, leveraged by Optimized Power Mode 3. Cap SoC interconnect frequency at 2.2GHz	1. Disable Perf P-Limit – Saves up to ~14W Power on average on Idle Socket [1] • Decouple uncore frequency selection on different sockets - save power or improve performance while another socket idles or runs a different kind of workload • Enabled by default in EMR and SPR, disabled in OPM 2.0 2. Enhanced Active idle Mode – Improves Performance in Core-bound Applications • Included C0 residency in Active Idle Mode calculation • Allows for Uncore frequency to drop below 1.4GHz, by avoiding false AIM detection • Available OOB with UP=0, leveraged by Optimized Power Mode 2.0 and sets UP=3 3. Core Count Aware Active Idle Mode • Improve active idle mode entry/exit based on cores used • Dynamic utilization point threshold setting • Available OOB, leveraged by OPM 2.0 4. FastC1e and SoC Interconnect Frequency cap at 2.2Ghz (Same as Optimized Power Mode 1.0)

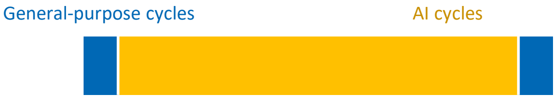
Intel Resource & Documentation Center: 817800

[1] OPM claims relative to EPB in Performance mode. Results may vary.

Understand your customers’ usage model

Large-Scale Dedicated AI

AI is the dominant workload

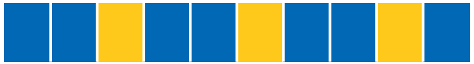


Dedicated AI application or service

Clusters based on
GPUs or AI accelerators

“General-Purpose” AI

AI is one of many workloads



Multi workloads (including AI) running
on the same infrastructure

Building and deploying at
enterprise scale on CPUs

Intel® Advanced Matrix Extensions (Intel® AMX) Acceleration Engine

What is Intel® AMX?

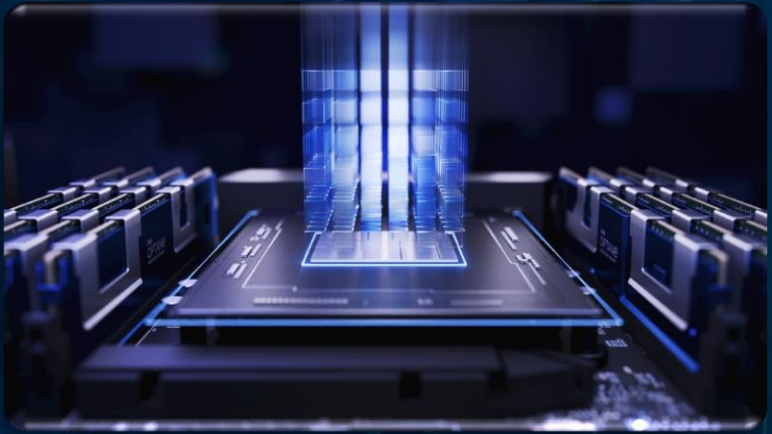
- Intel® AMX is a built-in accelerator that improves the performance of **deep learning** training and inference on 4th Gen Intel® Xeon® processors
- Advanced matrix multipliers are integrated into **EVERY** core

Business Value

- Help to **lower customers' TCO** as it raises the bar for where they can meet AI SLAs without the need for a discrete accelerator

Software Support

- Works **out-of-box** on industry-standard frameworks, toolkits and libraries such as PyTorch, TensorFlow, and OpenVINO
- **vSphere 8 supports Intel AMX**



PyTorch Training and Inference

Up to

10x higher

PyTorch for both real-time inference and training performance with built-in Intel AMX (BF16) on 4th Gen Intel® Xeon® Scalable processors vs. the prior generation (FP32)

See [A16, A17] at intel.com/processorclaims: 4th Gen Intel® Xeon® Scalable processors. Results may vary.

Complete Your Session Evaluations

Complete a minimum of 4 session surveys and the Overall Event Survey to be entered in a drawing to **win 1 of 5 full conference passes** to Cisco Live 2025.

Earn 100 points per survey completed and compete on the Cisco Live Challenge leaderboard.

Level up and earn **exclusive prizes!**

Complete your surveys in the **Cisco Live mobile app.**

Continue your education

- Visit the Cisco Showcase for related demos
- Book your one-on-one Meet the Engineer meeting
- Attend the interactive education with DevNet, Capture the Flag, and Walk-in Labs
- Visit the On-Demand Library for more sessions at www.CiscoLive.com/on-demand

The bridge to possible

Thank you

#CiscoLive