OCP SUMMIT

March 20-21
2018
San Jose, CA

OPEN. FOR BUSINESS.

Power Capping and Scheduling on Racks with Flexible Power Supply

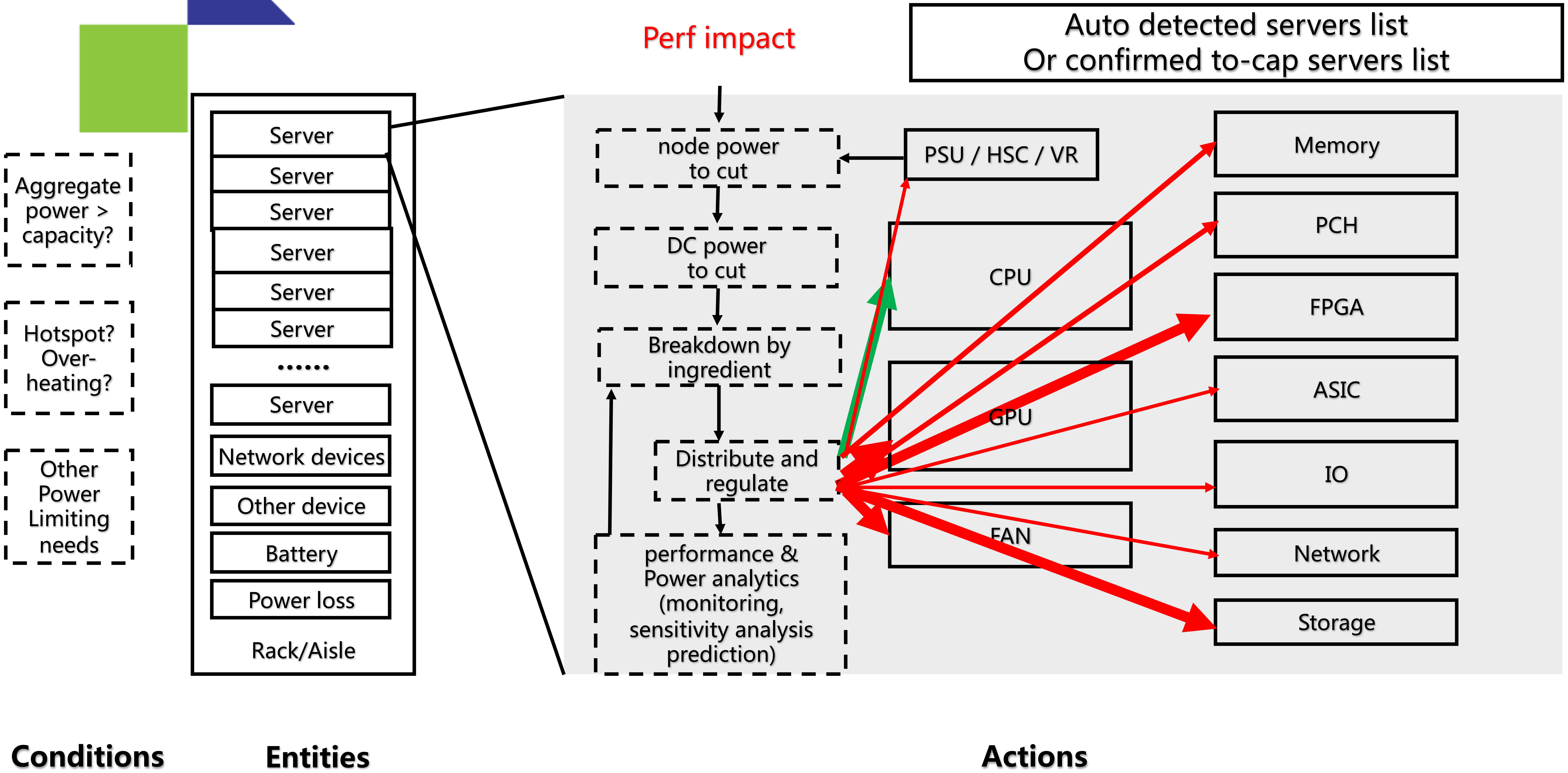
Justin Song, Chief Power Architect, Alibaba Cloud

Agenda

- Rack power capping
- Scheduling on racks with flexible power supply
- Summary & recommendation

Rack Power Capping

green : h/w or f/w ; red : s/w

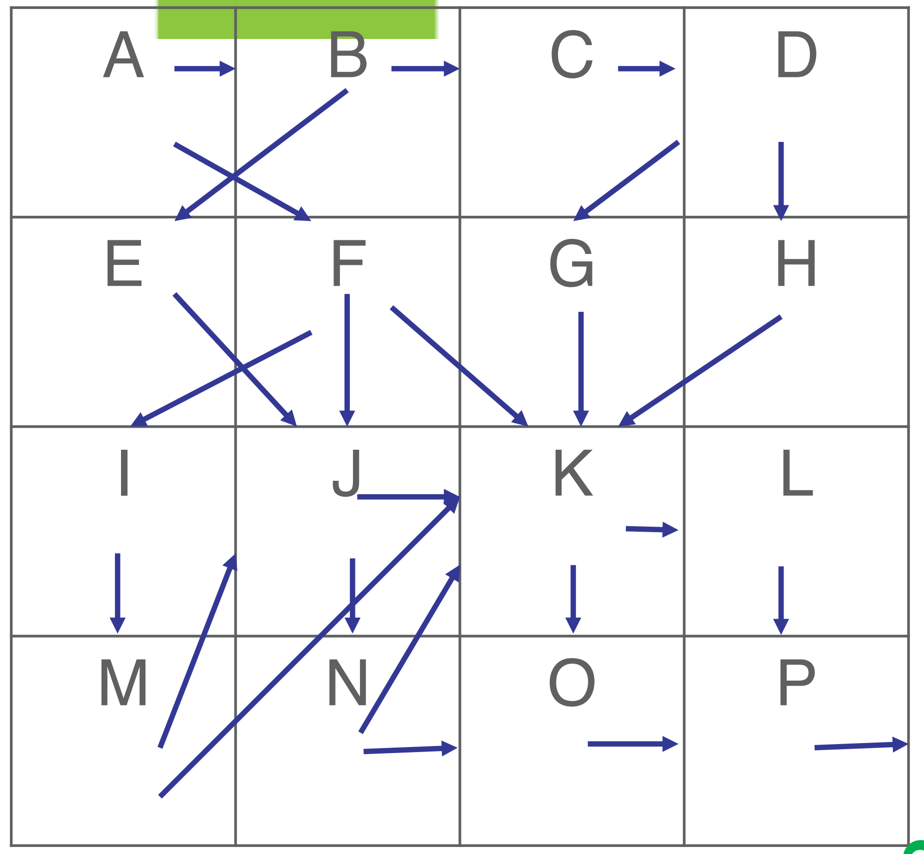


Top 5 reasons

- Save energy & costs
- DC power limiting
- Power exceeds rack quota
- Cooling fault
- Power supply failure

Capping Helps and Hurts

16 instances and dependency



Mapped to 12 nodes @ 4 racks

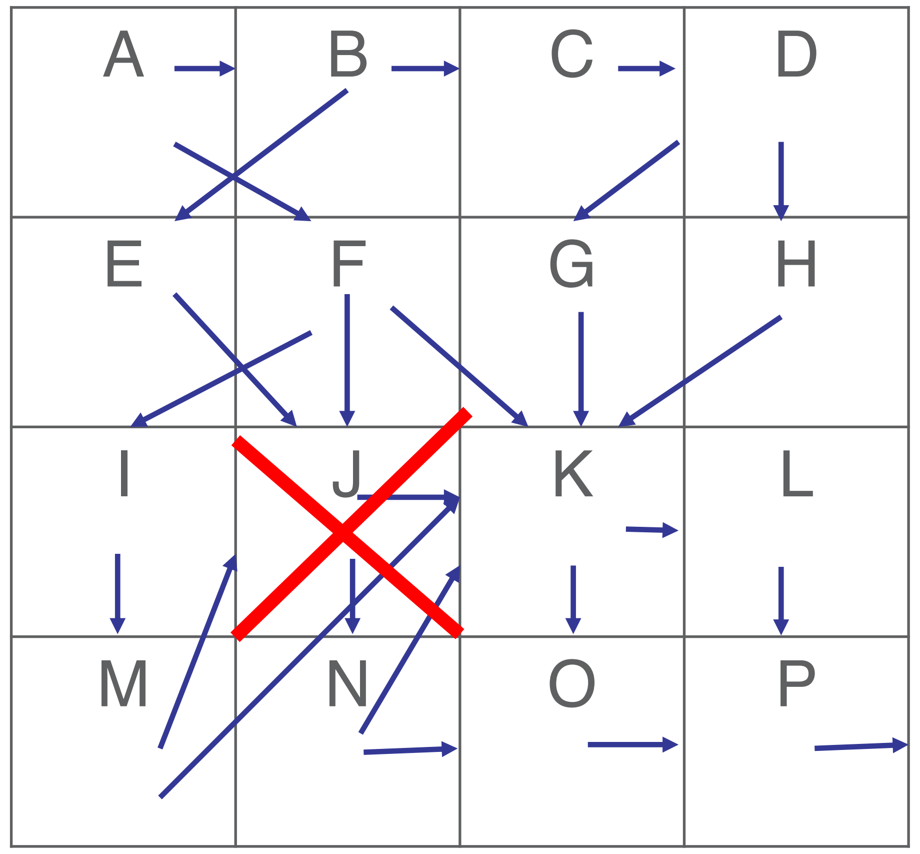
- J @ rack_3.node_5, 2.5GHz+
- K @ rack_4.node_7, 2.5GHz+
- A through I @ rack_11.node_6 through 10, 2.5GHz
- L through P @ rack_15.node_11 through 15, 2.5GHz

Capping = ON

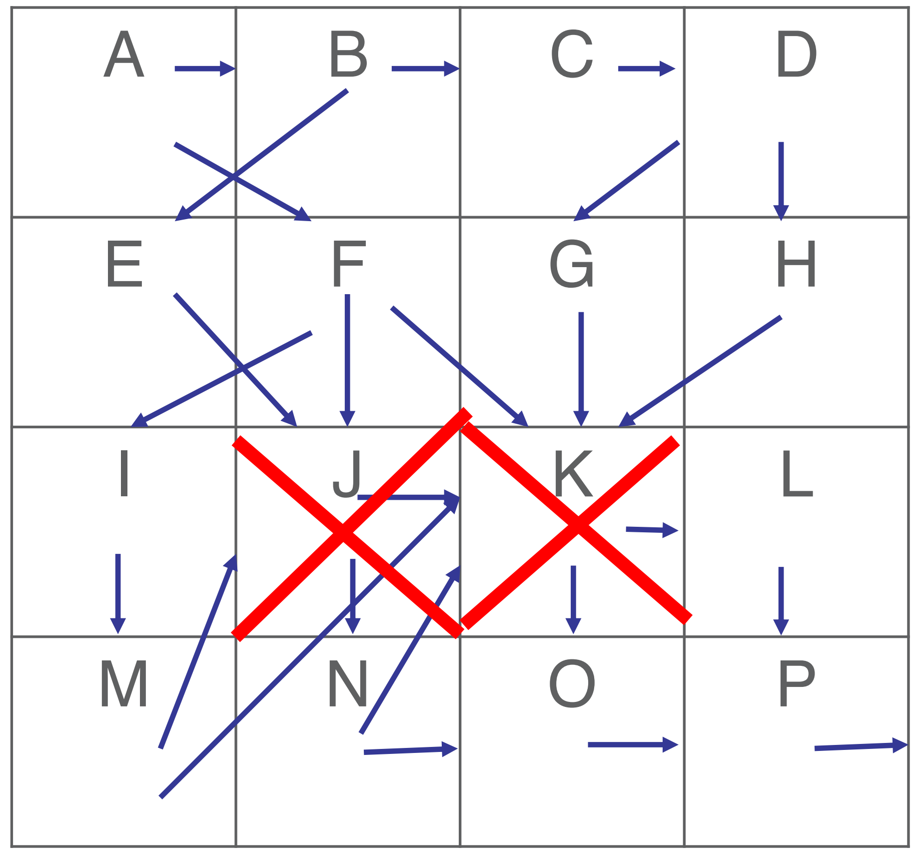
Power capping happens

- Rack 3 in failsafe mode (S5)
- Rack 4 @ 1.2GHz
- Rack 11 @ 2.0GHz
- L through P @ rack_15.node_11 through 15 @ 2.0GHz

Consequence T



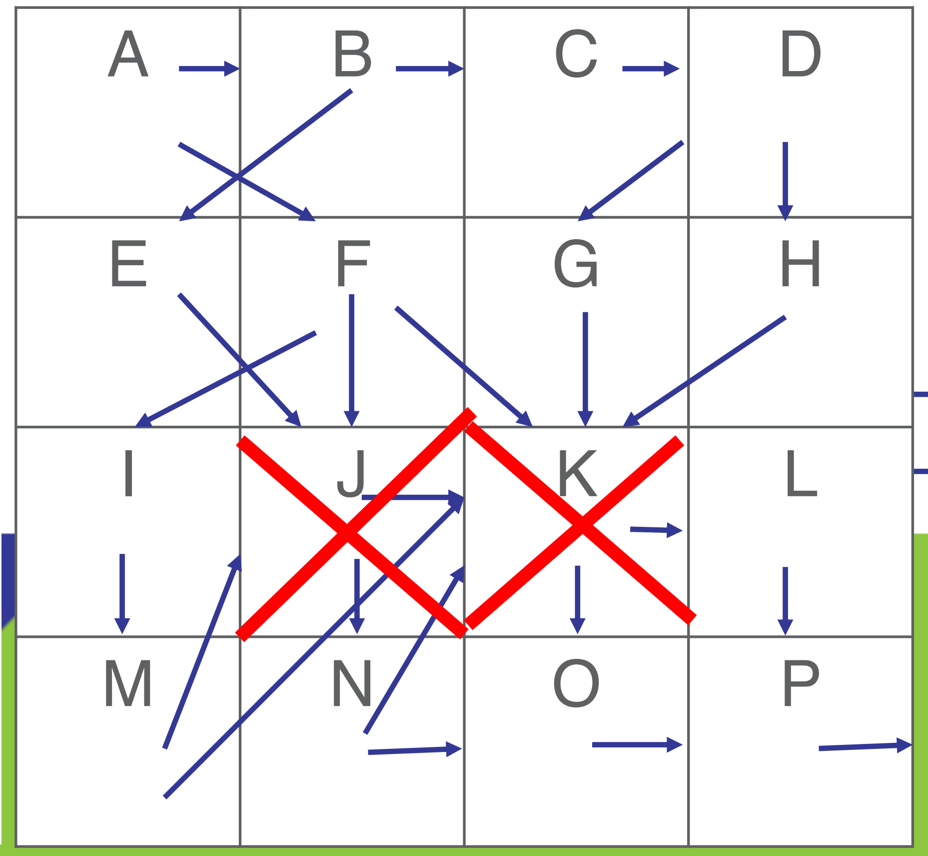
Consequence T+1



K's buffer blown

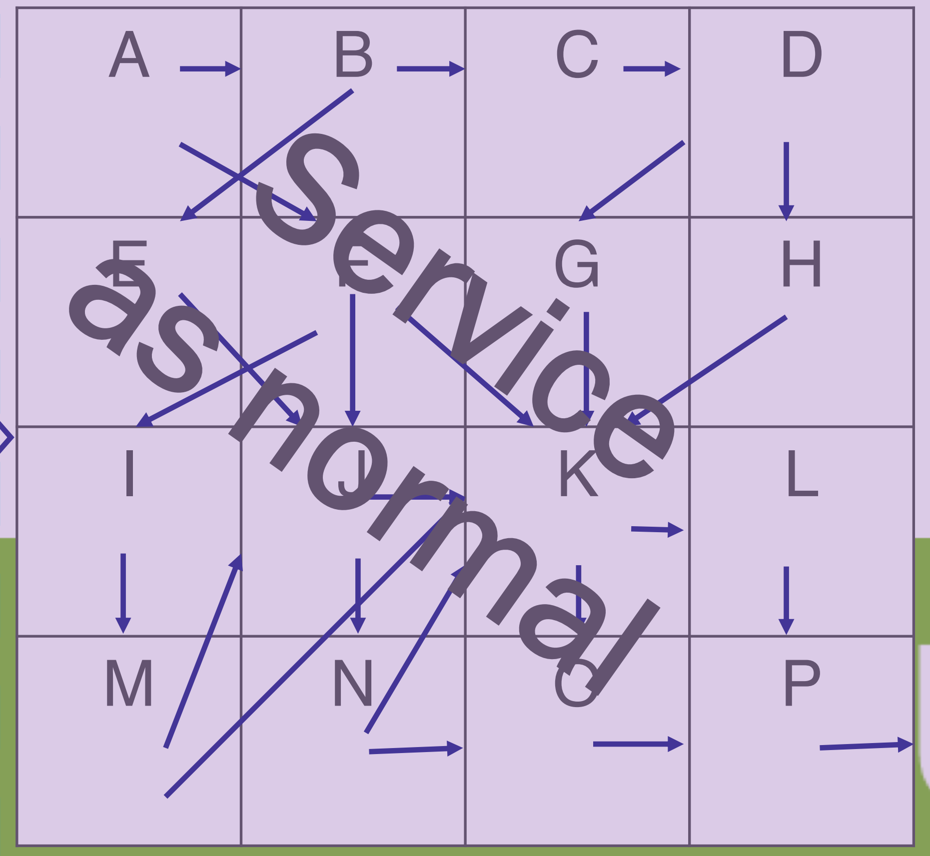
Capping = OFF

Consequence T+1

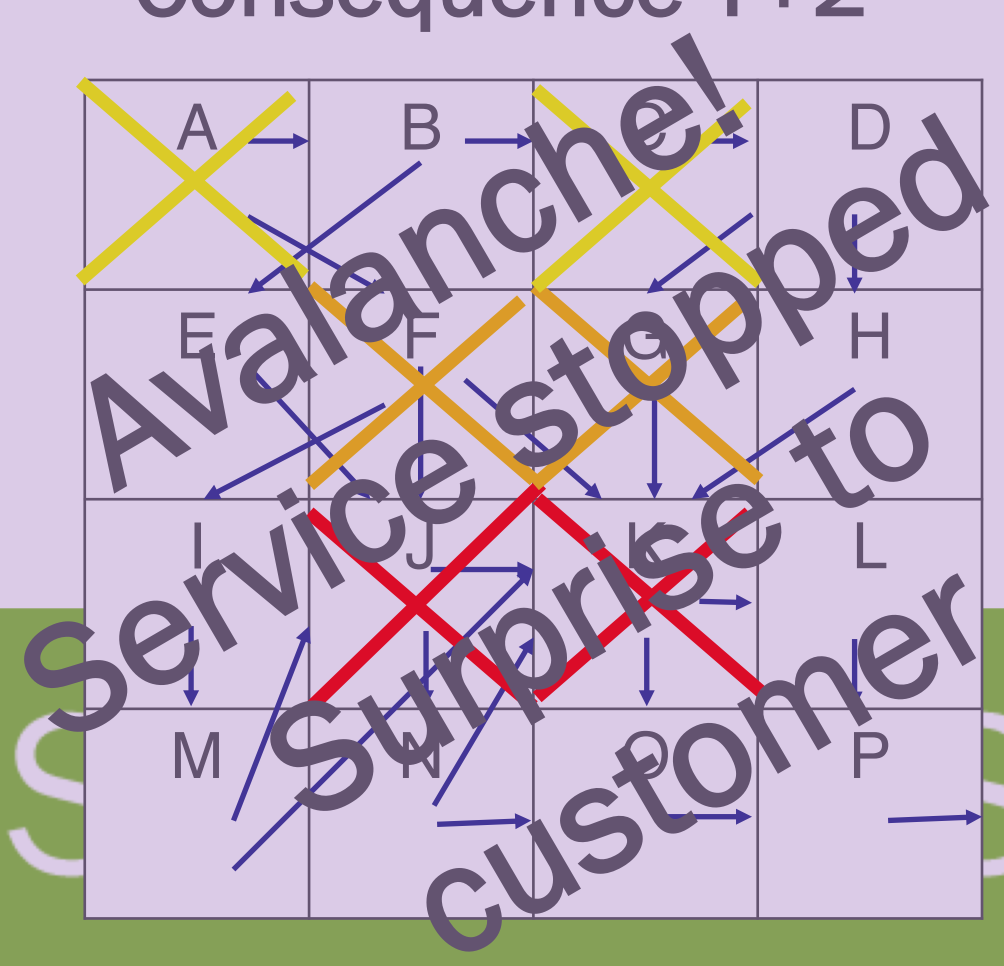


Pause, shutdown, migrate and resume

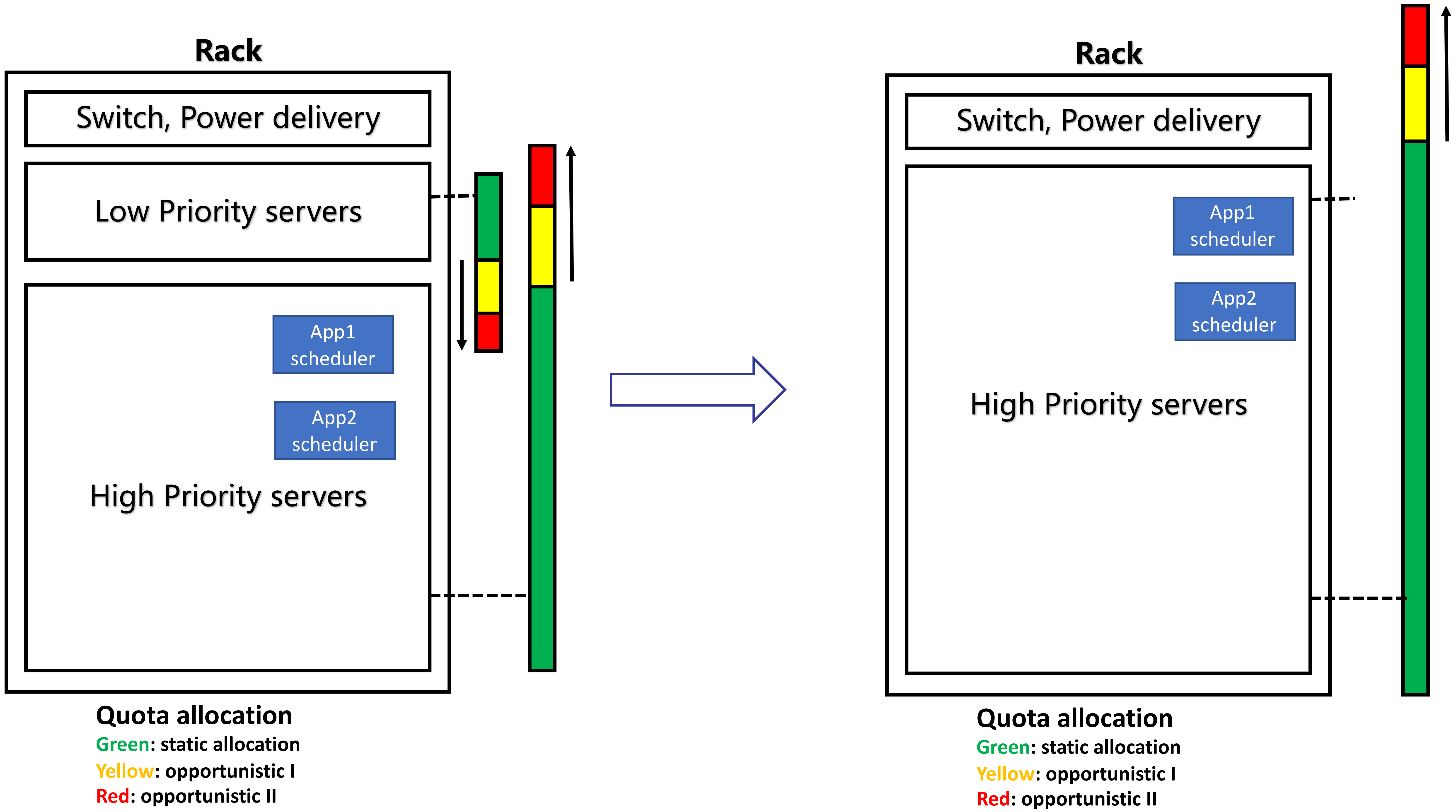
Consequence T+2



Consequence T+2

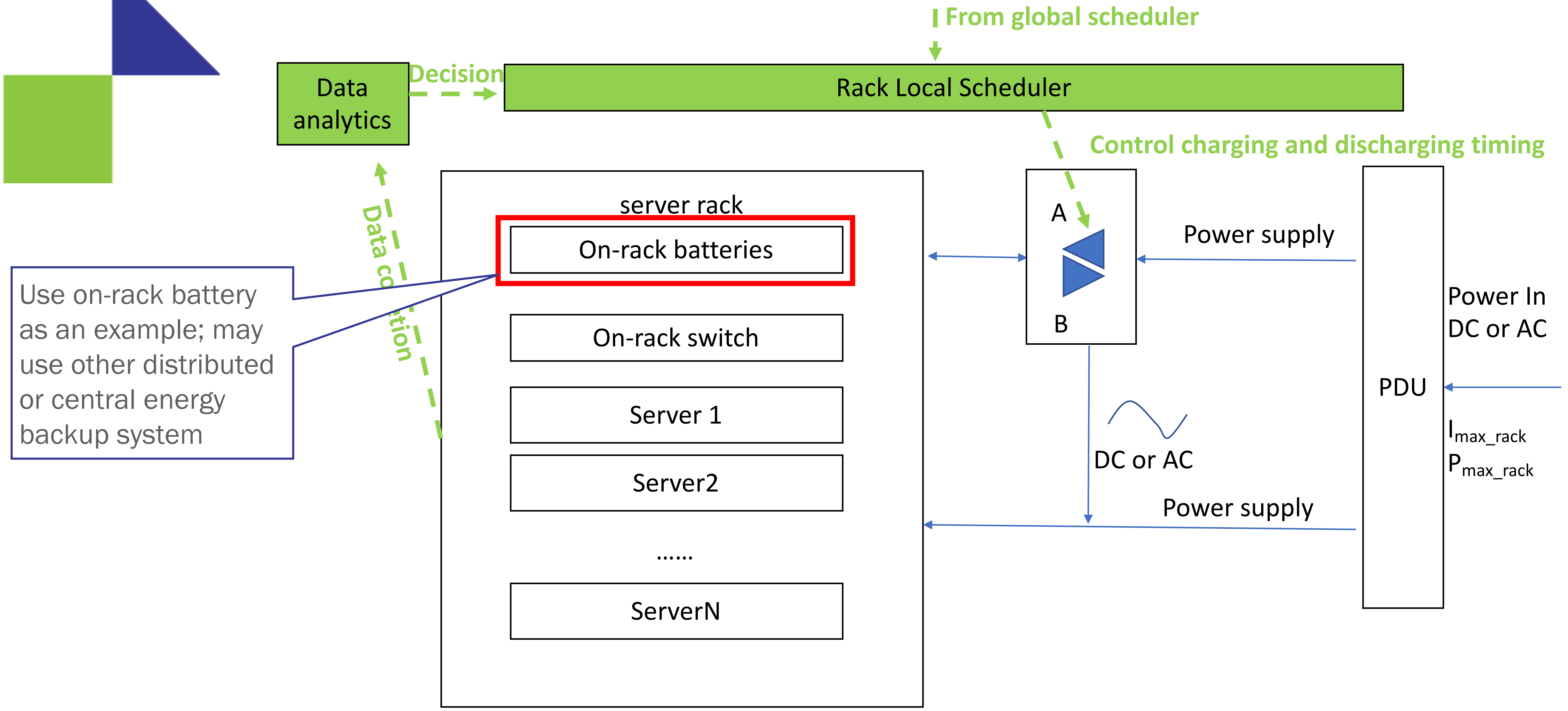


Another Challenge: Apps Deployment



We need to increase density and simply apps deployment without compromising SLA

Scheduling with Flexible Power Supply



Scheduler + Data Analytics → Use Battery Smartly

Manage Timing

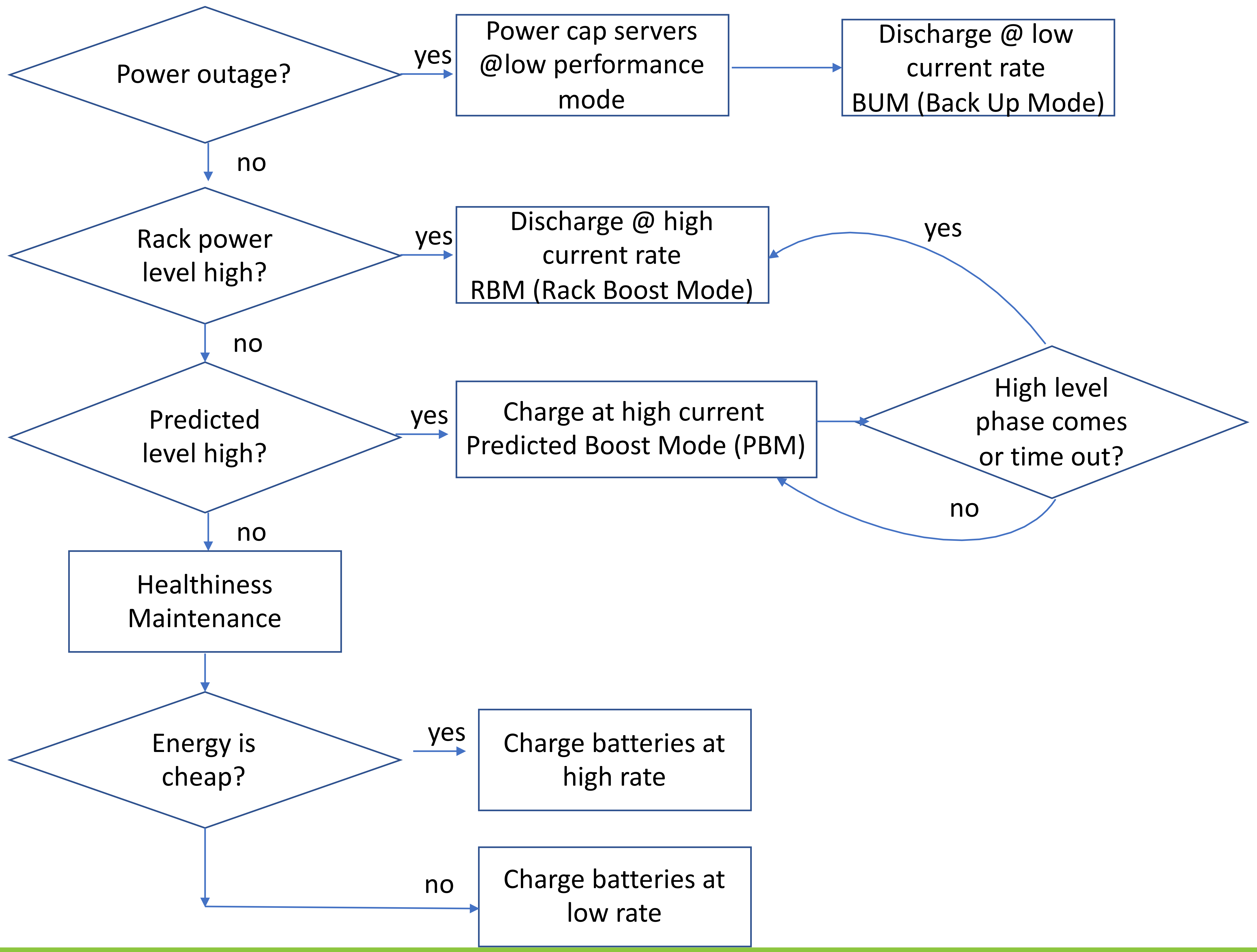
Charging

- Avoid peak usage of server loads
- Charge when energy is cheap

Discharging

- To accommodate max usage (Rack Turbo)
- Use as a UPS (power shortage or outage)
- Predict duration of high demand

Algorithm


Power outage?
yes
Power cap servers @low performance mode
Discharge @ low current rate BUM (Back Up Mode)
no
Rack power level high?
yes
Discharge @ high current rate RBM (Rack Boost Mode)
yes
no
Predicted level high?
yes
Charge at high current Predicted Boost Mode (PBM)
High level phase comes or time out?
no
no
Healthiness Maintenance
Energy is cheap?
yes
Charge batteries at high rate
no
Charge batteries at low rate

Summary and Recommendation

Use backup energy source to boost performance or enable higher deployment density

Nimble, flexible and cost effective deployment in data center

Software coordination, scheduling and application awareness is key

Intel Rack Power Optimization Technology

Nishi Ahuja, Principal Engineer, Intel

Rack Density Decreasing vs. Rack Power Under-utilization

Rack Density is limited by Power

Life Cycle

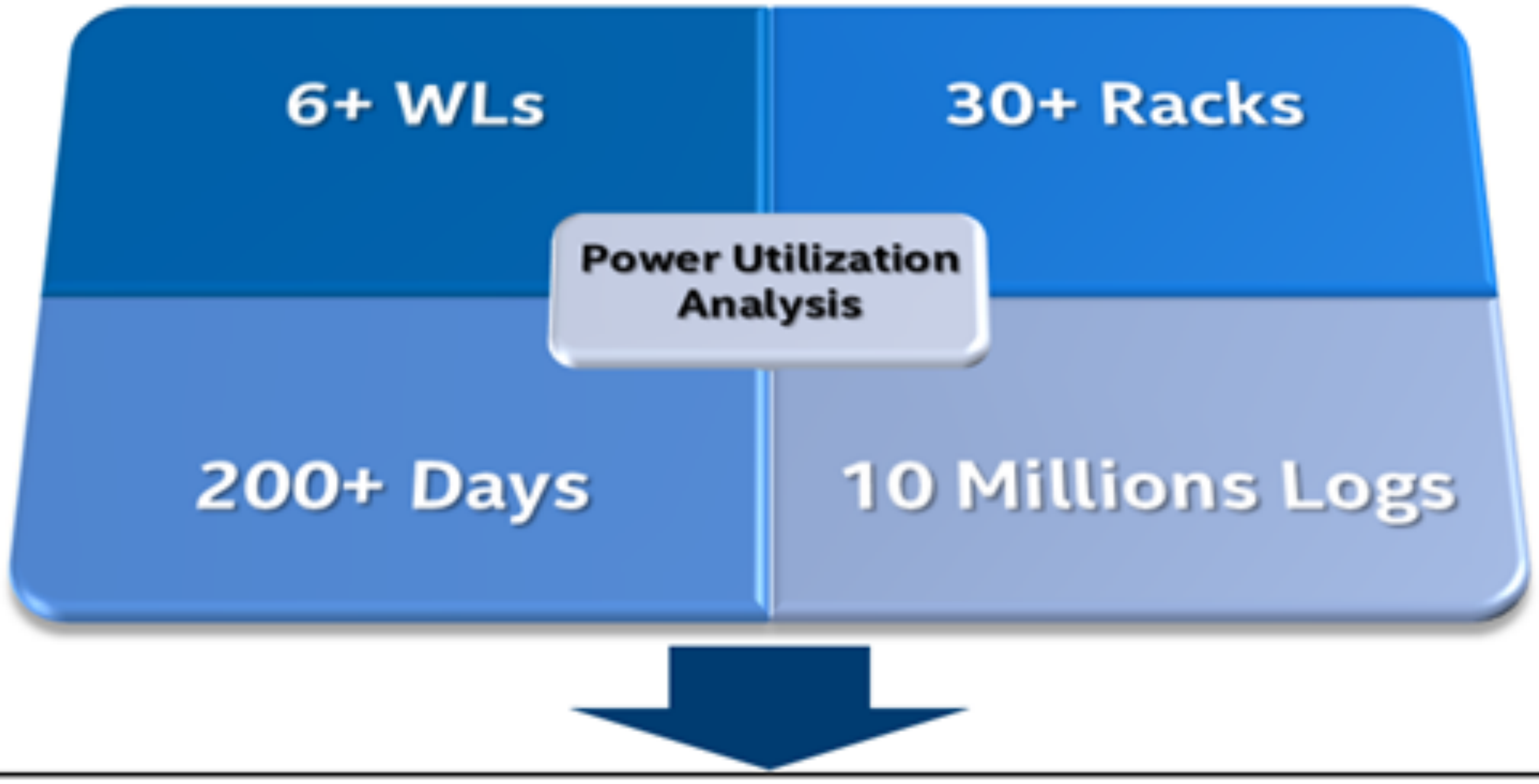
- Rack lifecycle is 10~15 years, server 3~5 years in datacenter
- **3 Server generation per rack generation**

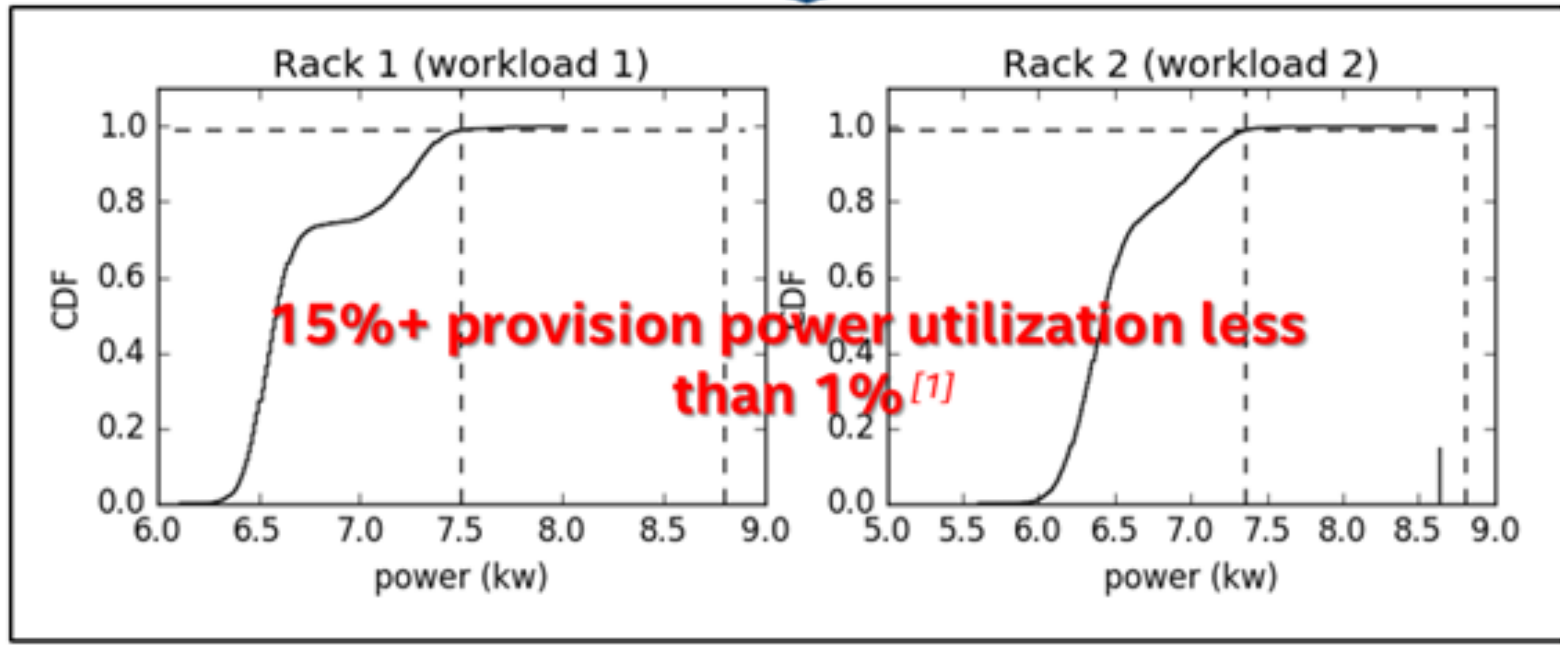
Stalled

- Rack Power Density can't be easily increased and **stalled at design target** (in PRC cloud datacenter, mostly 5-7KW)

Critical

- Server Node Density in Rack is **important TCO factor**





15%+ provision power utilization less than 1% [1]

<u>Problem Statement</u> – Rack density kept decreasing due to unmatched refresh cycle betweem infrastructure and IT equipment. The power usage analysis with top cloud service providers also showed the power utilization ratio is is lower than then planned due to dynamic range of peak to average.

[1] The evaluation is based on specific CSP workload and Intel Xeon Processors.

Node Manager Telemetry on Intel Xeon Scalable Processor

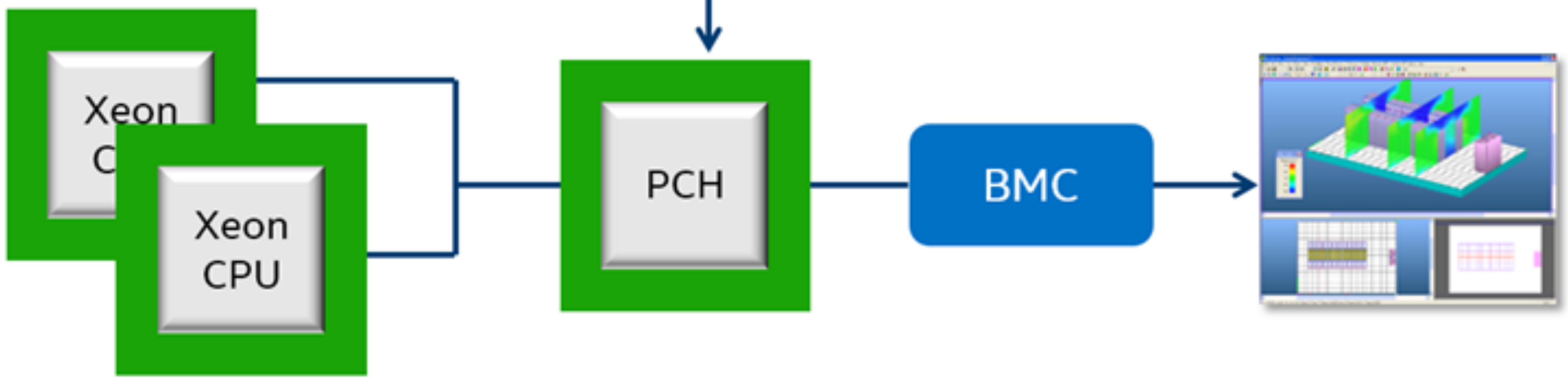
Power Telemetry
- Total platform power
- Individual CPU, Memory and Xeon Phi power domains

Thermal Telemetry
- Inlet & Outlet Airflow temperature
- Volumetric Airflow

Utilization Telemetry
- Aggregate Compute Utilization per sec
- CPU, Memory and I/O Utilization Metrics

Power Controls
- Power limiting during normal operation
- Power limit during boot

Intel® Node Manager
Firmware embedded in PCH's

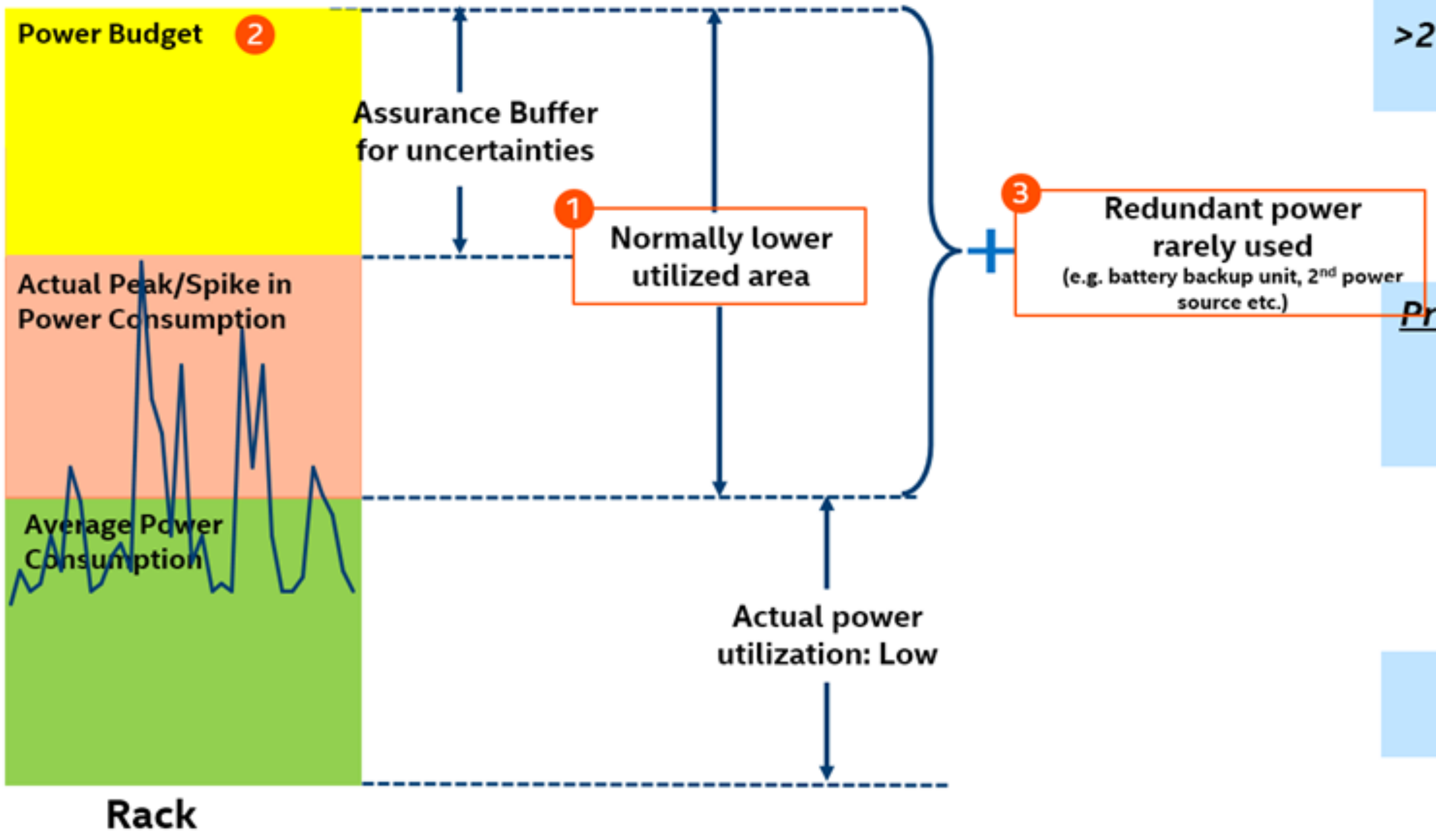


Intel Node Manger is widely adopted by Server OEMs and supports IPMI and DCMI API calls

Intel Rack Power Optimization Technology

Options	Description	Benefits
Static Power Capping	Configure power budget for each server node to avoid power violations	Rack Power Safety
Group Level Power Capping	Dynamiclly adjust rack power allocation among servers to enforce rack wide power consumption is within power budget line	Rack Density and Power Utilization
Dynamic Rack Power Provision (DRPP)	Leverage rack battery backup system to shave those sporadic power spikes or peaks over budge line	Rack Density and Service Reliability
Intelligent Orchestration – Power Awareness Scheduling	Schedule workload according to telemetry intelligence, e.g. power awareness job scheduling according to real power demand as well as available power capacity in rack (or cluster)	Energy Efficiency and Service Reliability
Dynamic Cluster Performance management	Dynamically adjust platform performance states according to rack power demands and actual resource utilization.	Performance Per Watt Efficiency

DRPP - Dynamic Rack Power Provision



>15%[1] of provision power, the utilization ratio *less than 1%*.
>20% of power capability(BBU), locked as back-up and rarely used[2].
Total >30% of power capacity is under-utilization!

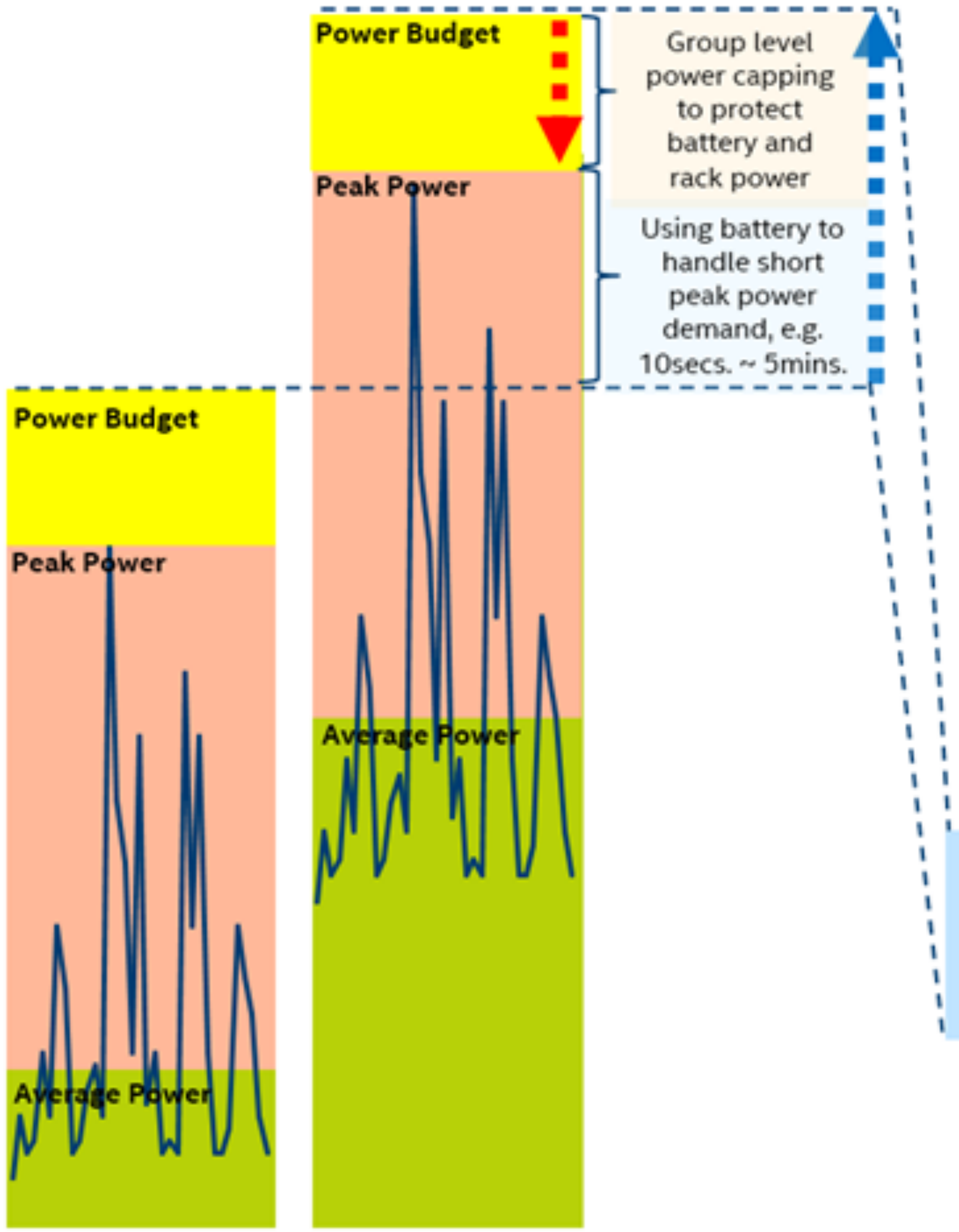
Proposed Solution Goal:

1. Decrease **power over-provisioning margin**
2. Dynamically increase **peak capacity**
3. Increase utilization of **back-up** power

Intel Dynamic Rack Power Provision Technology
to balance power utilization and improve rack density

[1] – 15% is concluded from analysis of typical online rack system, and subject to change with environment change.
[2] - The general design of backup power unit capacity is same as rack power capacity, and 20%~30% of backup power capacity is rarely used even under power failure condition.

Turbo Rack - Innovative, Non-Disruptive Solution



Software - Power Awareness Intelligence
- Real-time Power Insights
- Group Power Capping and Dynamic Peak Capacity Provision
- Intelligent Job Scheduling/Migration

Application Interface - RESTful
- Contributed as Redfish API
- Easy Application Integration

Hardware– Distributed Energy System
- Open Interfaces for 3rd party battery system integration
- Unified architecture support, including OCP and conventional rack

Firmware – Algorithm Optimization
- Adaptive Power Re-balance Algorithm
- Configurable Power Limiting

Enabled by ***Dynamic Rack Power Provision Technology*** with improvement on power utilization and rack density ***by 15%~25%*** [1]

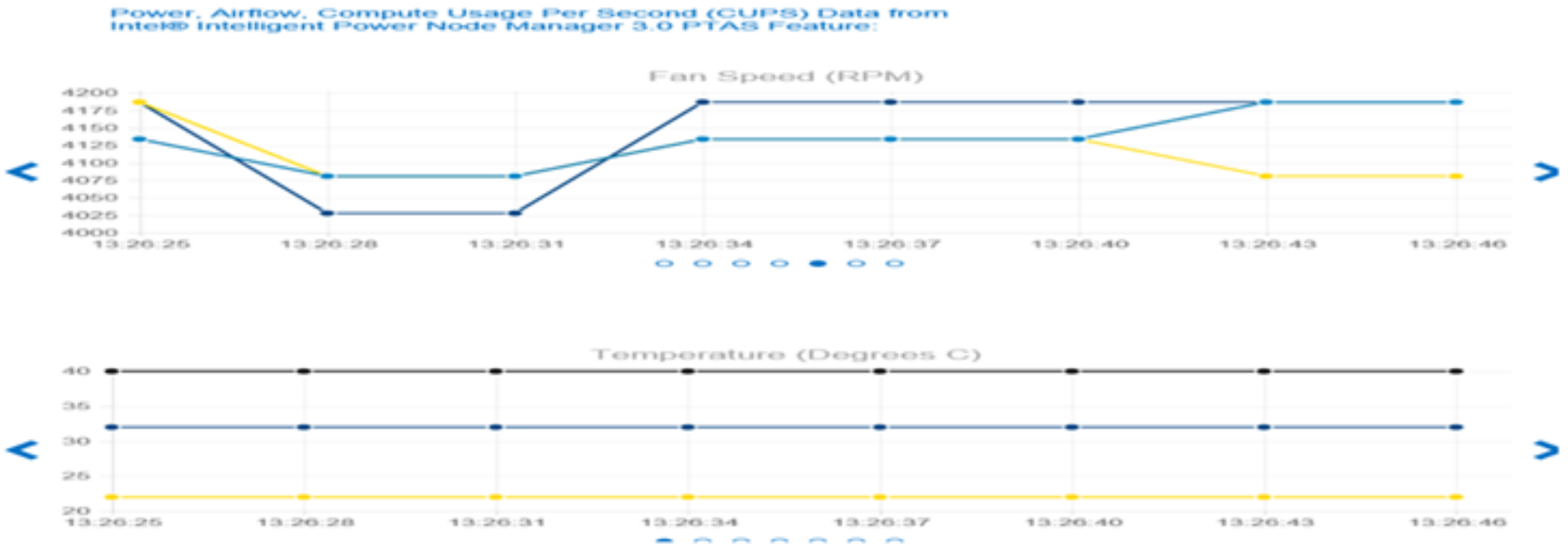
Power Shelf for OCP Rack [2]

Distributed Battery Backup System

[1] – 15%~25% is per evaluation analysis from specific workloads(WLs), and subject to change for different environments.
[2] – Refer to https://datacenterfrontier.com/ocp-compute-rack-makers/

Intelligent Orchestration based on HW Telemetry & Analytic

Strategy	Descriptions	Telemetry and Analytic
Power Awareness Scheduling	Schedule/Migrate workload according to power intelligence, e.g. energy efficiency, power budget, dynamic power capacity	System/rack power, power analysis.
Uniform airflow/Thermal Condition scheduling	Dynamic adjust workload placement to get one uniform thermal condition to avoid hot-spot in datacenter (with better PUE as well as)	Inlet/outlet temperature, Return Temperature Index(RTI), Rack Cooling Index(RCI)
Failure Event Triggered Migration	Migrate workload to other cluster or rack in case some critical failure events (e.g. power, or thermal) is identified	Failure events



Turbo Rack Video (Intel)

2minutes 30 seconds

OCP
SUMMIT