Real Time Monitoring and Availability of Platform Telemetry for Efficient Data Center Cooling

Intel Corporation
Data Center Platform Application Engineering

May 2013
Reference Number: 523447
Revision Number: 1.0

Agenda

- Problem Statement
- Methodology for Addressing Cooling in-efficiencies with PTAS
- Platform Enabling details
- Summary and Call to Action

Agenda

- Problem Statement
  - Data Center challenges
  - Data Center Power consumption
  - Data Center cooling in-efficiencies
- Methodology for Addressing Cooling in-efficiencies with PTAS
- Platform Enabling details
- Summary and Call to Action

Data Center Challenges

- Business growth is leading to significant demands on IT & Facilities
- Profitability depends on efficient capital investment and operational efficiency

Dynamic Market Environment (2015)

More Users
>3 Billion Connected users[1]

More Devices
>15 Billion Connected Devices[2]

More Data
>1.5 Zetabyte Of cloud Traffic[1]

- Total power consumed by Data Centers ..2-3% of all electricity generated by 2014..EPA
- $27 B/yr spent on server energy costs..IDC 2009
- Data will grow 44 times to 35ZB between 2009 – 2020..IDC 2011

Business Growth

- **Significant "Capacity" demand**
- **Disruptive application & infrastructure trends**

Profitability - Dependencies

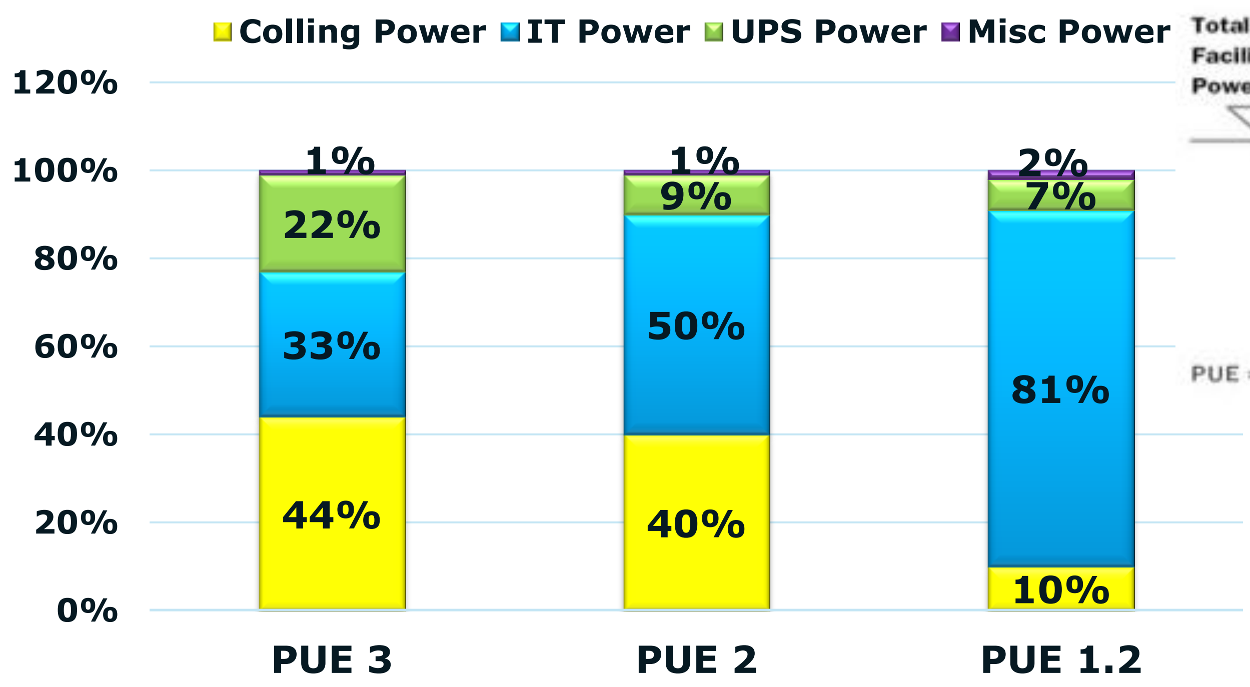
- **Capital Efficiency**
- **Operational Efficiency**
- **Capacity on Demand**

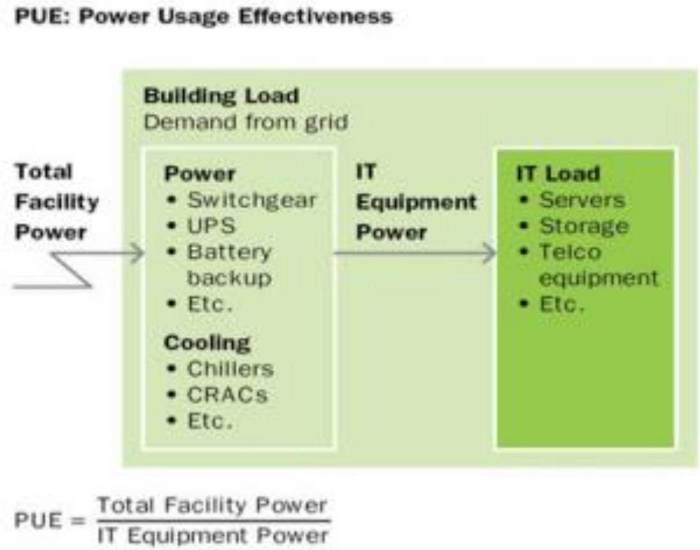
Business – Growth & profitability Opportunity

1. Cisco Global Cloud Index Nov 2011
2. Intel ECG "Worldwide Device Estimates Year 2020 - Intel One Smart Network Work" forecast

Data Center Power Consumption

Assumption - 1,100 Racks, 95% Populated racks, 44,000 Servers, 88,000 Intel® Xeon® Processors – Intel Internal study

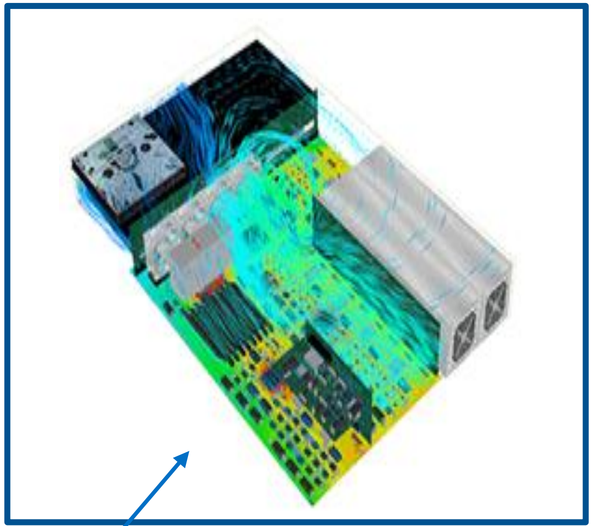




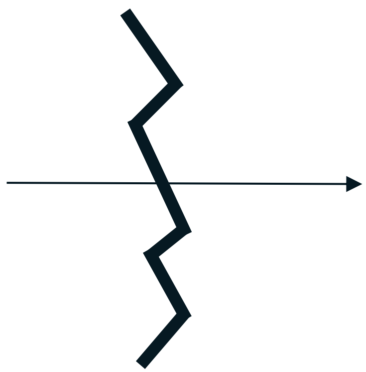
Optimize cooling power is critical to drive overall DC power efficiency

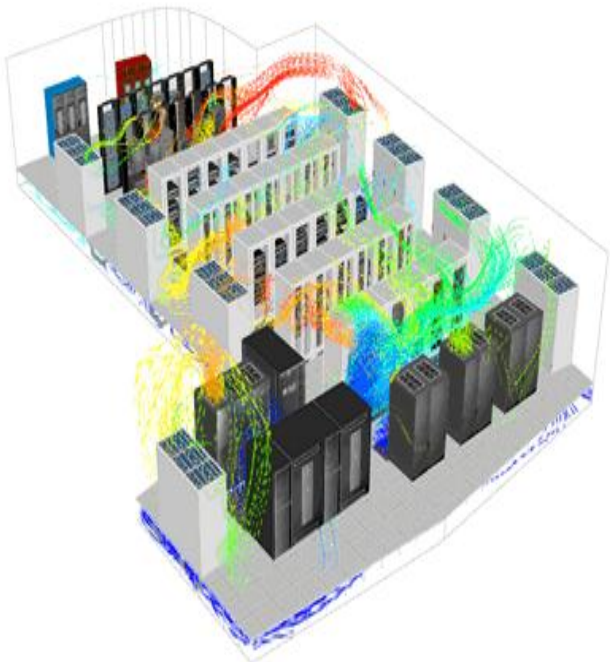
Current State of Data Center Cooling Control: The Server/Facility Disconnect

IT Equipment Manufacturer | **Facilities Manager**



Disconnect





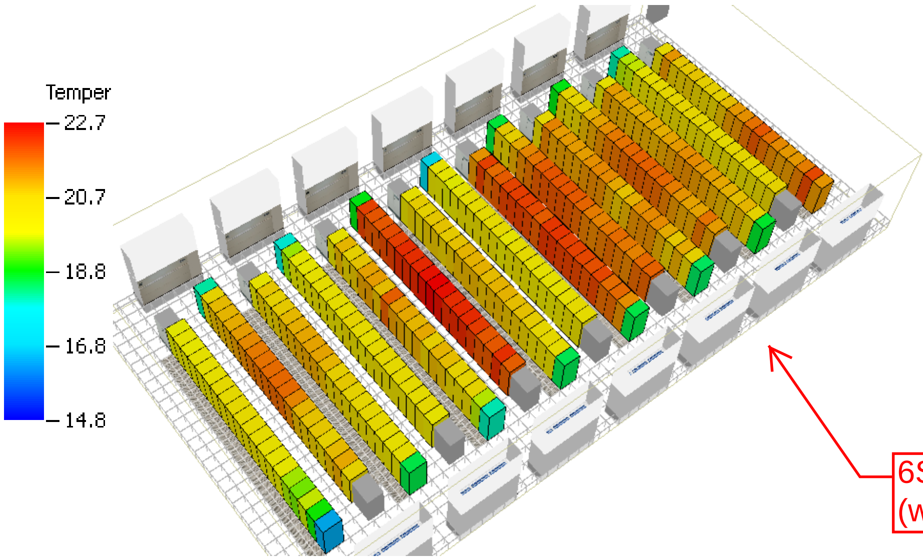
Platform Telemetry:
- Volumetric airflow
- Exit Air Temperature
- CUPs

6SigmaET
(www.sheaf.com.tw)

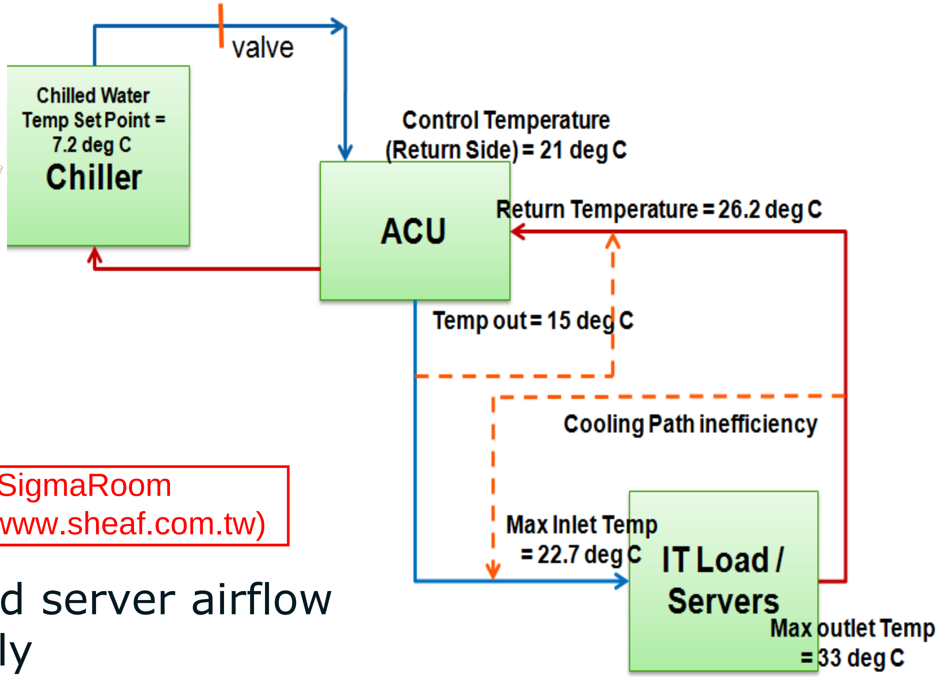
The facility and the servers are designed separately – a major cause of performance problems and low efficiency

1. SEMI-THERM, 2011 27th Annual IEEE, Ahuja, N.; Rego, C.; Ahuja, S.; Warner, M.; Docca, A.; "Data Center Efficiency with Higher Ambient Temperatures and Optimized Cooling Control"

Data Center Cooling in-efficiencies



6SigmaRoom (www.sheaf.com.tw)



- No good solution to match aggregated server airflow demand with CRAC/ACU airflow supply
- By-pass (excess of supply air)
- Re-circulation lead to data center hot spots
- Typically, data center cooling devices use return air temperature sensors as the primary control-variable

Inefficiencies in cooling path management must be understood and eliminated to achieve substantial saving

Evolution

Class and Upper Temperature Limit Recommended by ASHRAE				
Recommended	Allowable			
All 'A' Classes	A1	A2	A3	A4
18°C -27°C (81°F)	32 °C (90°F)	35 °C (95°F)	40 °C (104°F)	45 °C (113°F)

- ASHRAE recommended range, applies to server inlet conditions
- It is not applied to room temperature:
  - Either in hot and cold aisle
  - Under raised floor
  - CRAH return temperature

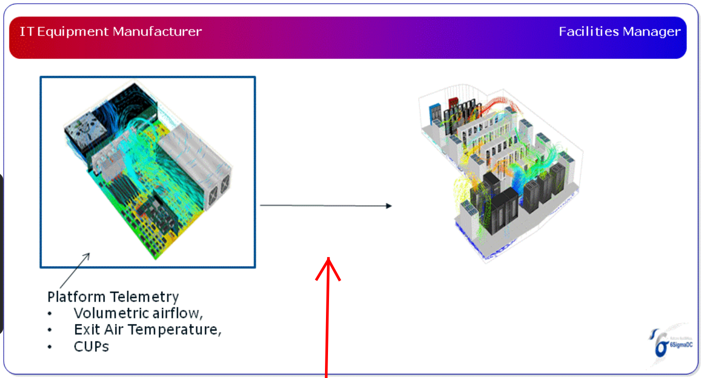
Control cooling on the supply side or Server/Rack inlet

1. 2012 ASHRAE Thermal Guidelines for Data Processing Environments

Data Center Efficiency

Cooling best practices

- Containment with integrated IT and Facility
  - Server manufacturer need to expose data
- Free-cooling with air or water economizers
  - Design for running without chillers when possible
- Expand data center temperatures and humidity ranges
  - ASHRAE defines new classes for DCs: A1 thru A4
  - A2 is typical (up to 35 ˚C IT inlet temperatures)
  - A3 is new (up to 45 ˚C IT inlet temperatures)
- Intel focusing on enabling the higher temperature data center
  - Platform design guide
  - Data Center design guide



6SigmaET+6SigmaRoom
(www.sheaf.com.tw)

Data Center efficiency – Drive to lower PUEs

1. Figure from SEMI-THERM, 2011 27th Annual IEEE, Ahuja, N.; Rego, C.; Ahuja, S.; Warner, M.; Docca, A.; "Data Center Efficiency with Higher Ambient Temperatures and Optimized Cooling Control"

Agenda

- Problem Statement
- Methodology for Addressing Cooling in-efficiencies with PTAS
  - Linking server telemetry to facility management software
  - Intel Solution - PTAS
  - The PTAS Approach
- Platform Enabling details
- Summary and Call to Action

Power Thermal Aware Solution (PTAS)

Intel Solution : Intel's Data Center Infrastructure Management (DCIM) solution with integrated platform telemetry and analytics to identify and address DC energy efficiency issues

PTAS Components

1. **Data (Platform telemetry)**
    1. PTAS CUPs
        1. Compute/Workload utilization
    2. PTAS Thermal
        1. Volumetric Airflow
        2. Outlet air temp
2. **Logics (Analytics)**
    1. OOB data collection
    2. Cooling and Compute Metrics
    3. Rules & Policies
    4. ACU 2 way communication & control
    5. Workload placement recommendation

PTAS Stack

1. Hardware layer (Grantley)
2. Firmware layer (Intel Node Manager 3.0)
3. Software layer (Intel DCM 4.0)

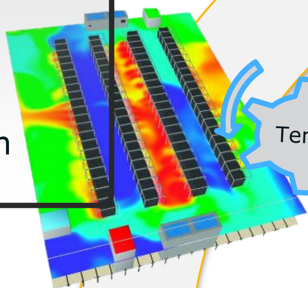
PTAS Ecosystem

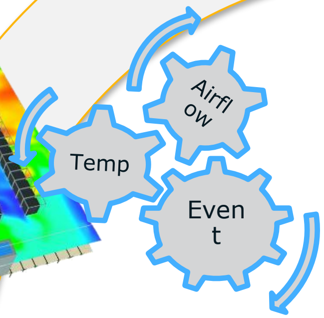
1. End users - Cloud Service Providers, Telcos, Hosters, Co-lo providers
2. OEM/ODMs
3. Independent software vendor (ISV)
4. Infrastructure Management Vendors (IMV)

PTAS Approach

3. Evaluation

Hot Spot
Cold Spot
Bypass
Recirculation





4. Manage Events

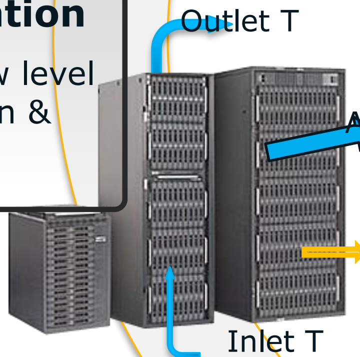
Cooling
- Adjust Temp
- Adjust Airflow

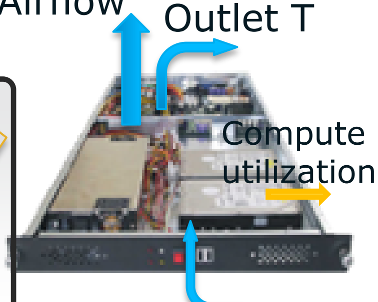
Return
- Uniform Temp
- Compute Energy

Mitigation
- Alerts
- Power policy

2. Aggregation

Rack and Row level data collection & aggregation





1. Collect Data

Platform Based Sensors at each server vs. 3rd Party Sensors in isolated locations

The Benefit

1. Opex savings
2. Cooling Capex savings
3. External sensor instrumentation savings
4. Extend Capital lifespan

Agenda

- Problem Statement
- Methodology for Addressing Cooling in-efficiencies with PTAS
- Platform Enabling details
  - PTAS Thermal
    - Enabling steps
    - PTAS Thermal Usage
    - Potential energy savings
  - PTAS CUPs
    - Validation steps
    - PTAS compute usage
- Summary and Call to Action

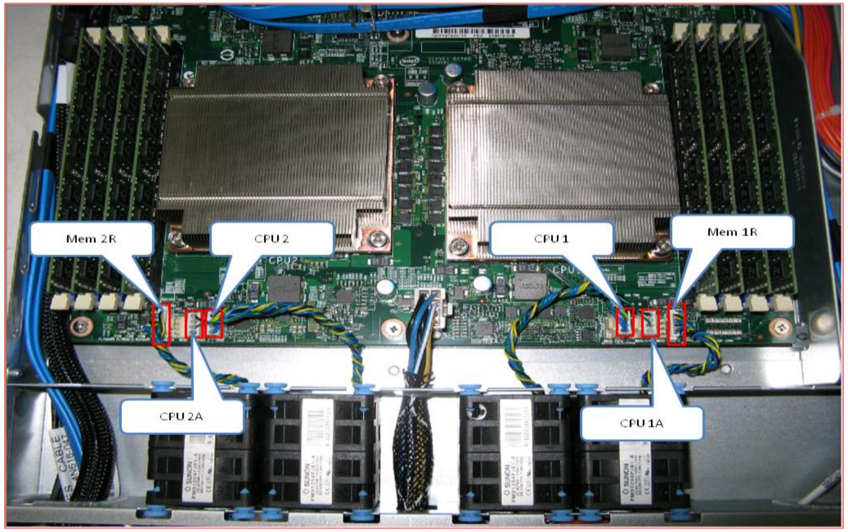
Availability of New Thermal Virtual Sensors

- The new sensors are defined for providing server level information
  - Total Airflow through the server
  - Average outlet temperature of server
- New sensors are derived from sensory data already available on the server
- The Airflow is derived from the speed (RPM) of each fan zone

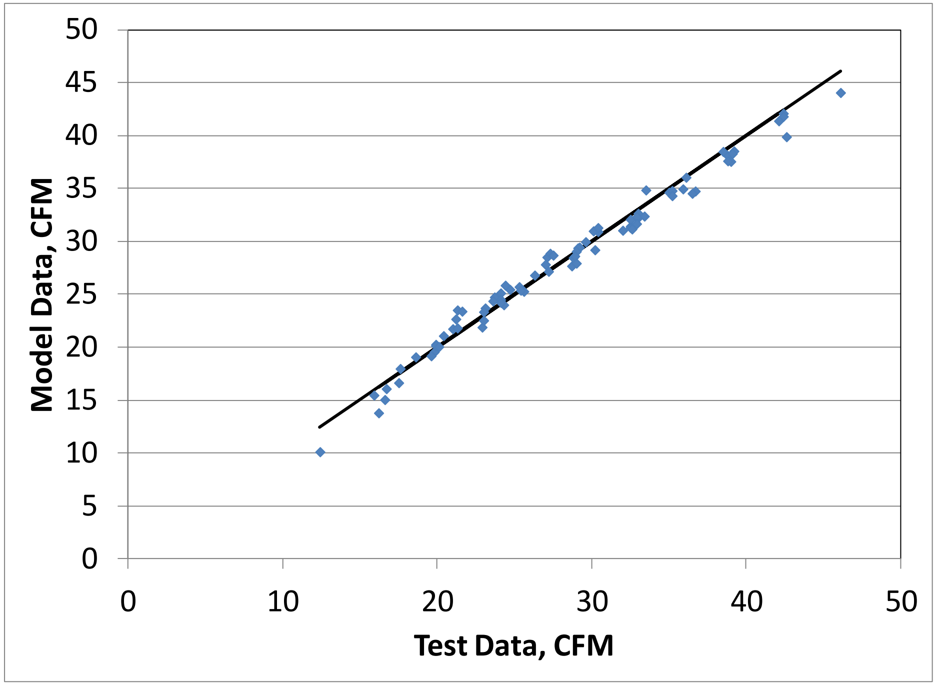
$$Q = f(RPM)$$

Where: Q = Server Airflow, RPM = Fan RPM (Available real time)

Method Accuracy



PWM 0		PWM1		PWM2		PWM3	
CPU Fan1	CPU Fan1A	Mem1R FanA	Mem1R FanB	CPU Fan2	CPU Fan2A	Mem2R FanA	Mem2R FanB



- Method has been validated several systems of different fan zones
- Coefficient of correlation (R2) of 0.981 indicating a very good correlation between volumetric airflow and the fan speeds
- Airflow can be computed in ME/firmware
- Computed airflow can be exposed as IPMI commands

Availability of New Thermal Virtual Sensors

- Outlet Temperature derived from Airflow, power dissipation, altitude, Inlet Temperature

$$T_{outlet} = T_{inlet} + \frac{1.76 \cdot P}{Q} \cdot f_{alt}$$

Where: T_{inlet} is the ambient temperature from the front panel sensor

Q is the volumetric airflow from the model based on sensed RPM values (New Derived sensor)

P is Exponentially averaged system power over a server thermal time constant ~100s

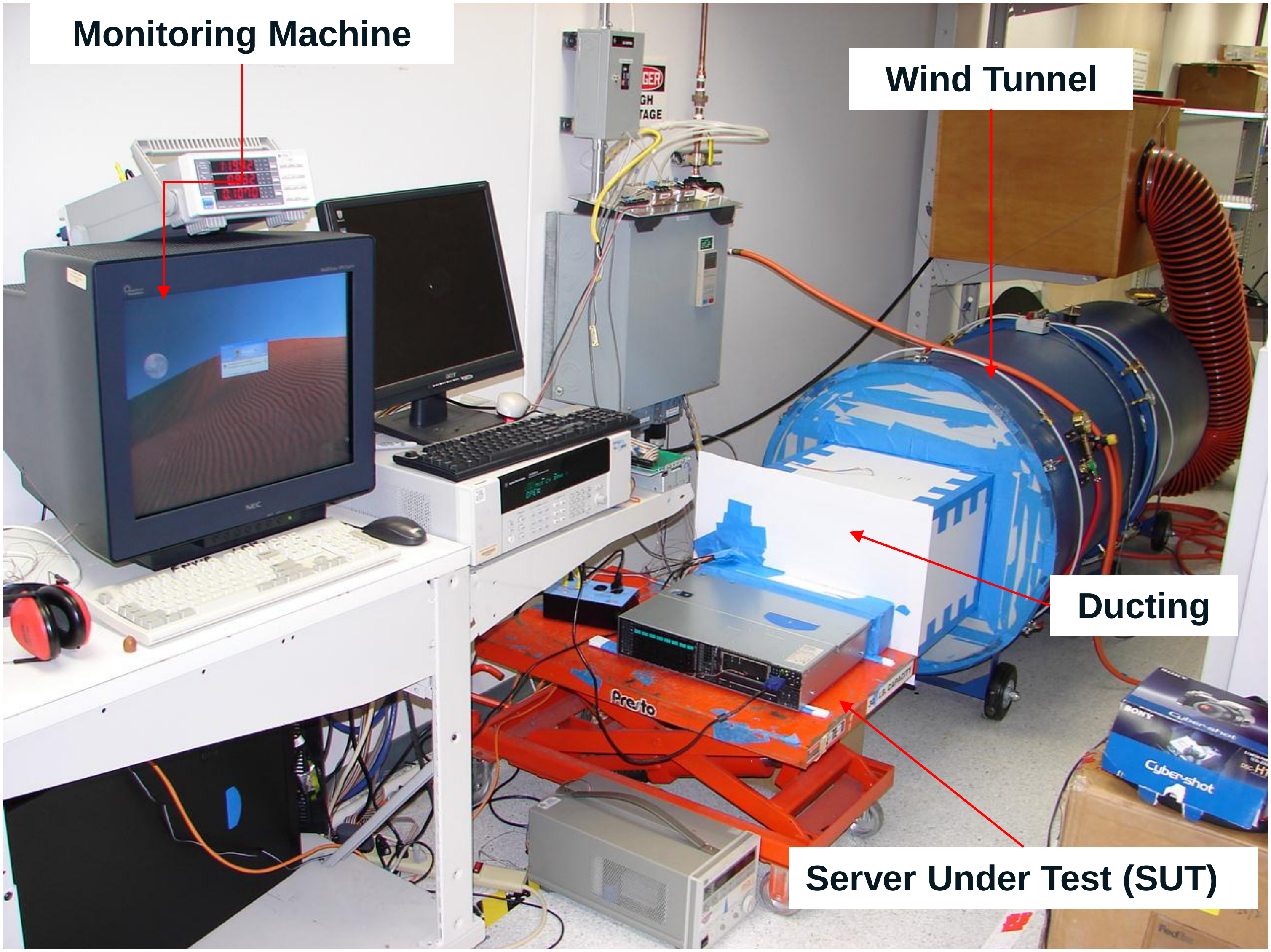
falt is the density correction factor

- Computed outlet temperature is exposed as Intel® Management Engine/BMC sensors

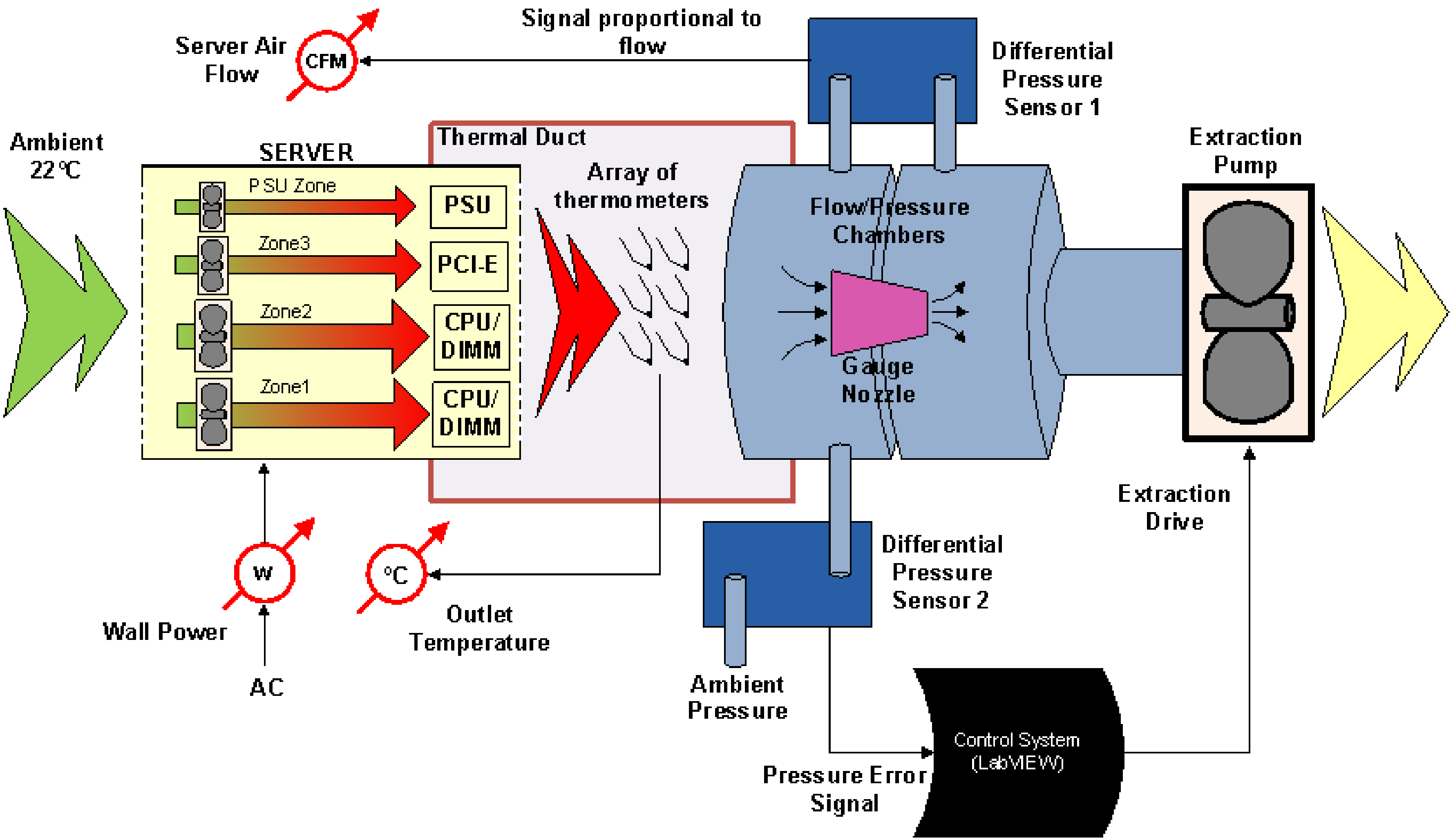
Thermal Platform Enabling Steps

- Set up and configuration
- Collect RPM values and CFM for given PWM
- Determine model coefficients
- Enter coefficients into appropriate Intel® Management Engine configuration file
- Expose volumetric airflow and outlet temperature (IPMI)

Airflow Characterization Setup

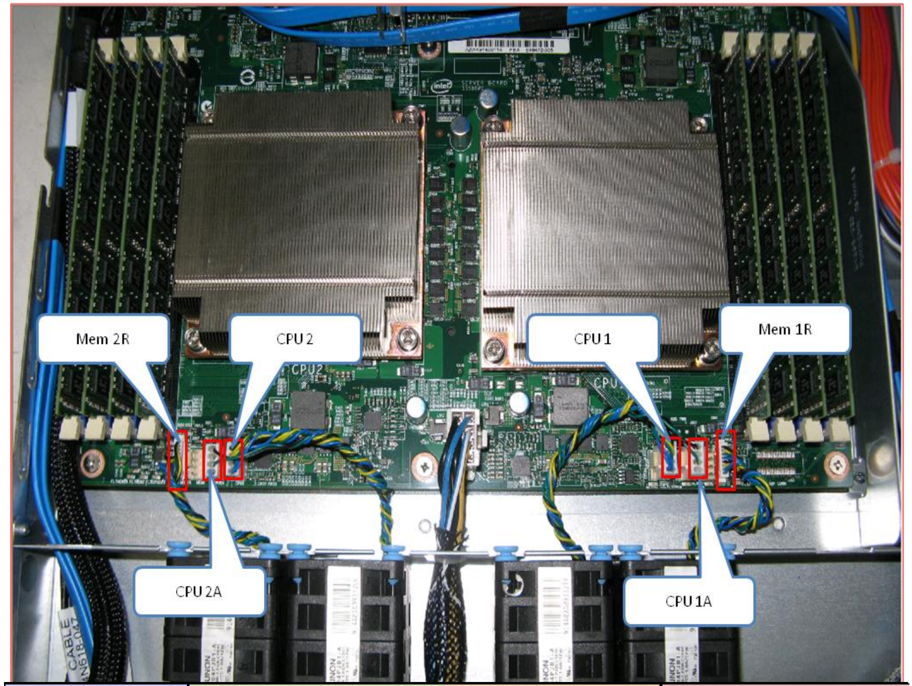


Experimental Setup - Schematic



Data Collection Methodology – Server with Four Fan Zones

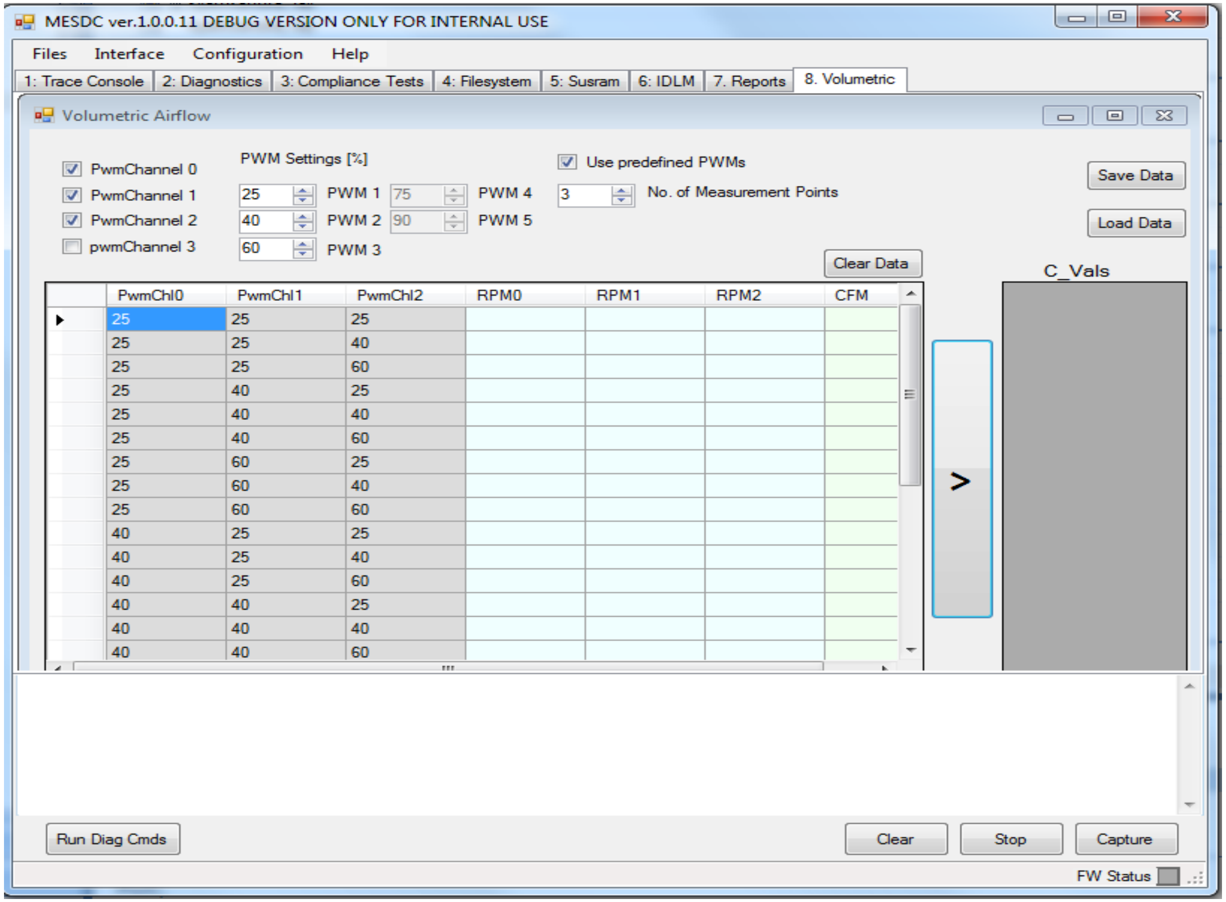
- Vary PWM for each zone from 25 to 60 (PWM settings: 25, 40, 60)
- Since there are four zones and three possible PWM values for each zone, there are 3X3X3X3=81 total possible combinations of PWM values
- Changing PWM values will result in varying RPM values. Record RPM values
- Measure airflow through the chassis using a wind tunnel
- Record 81 sets; each containing the following data
  - PWM1, PWM2, PWM3, RMP1, RPM2, RPM3, RPM4, CFM



PWM 0		PWM1		PWM2		PWM3	
CPU Fan1	CPU Fan1A	Mem1R FanA	Mem1R FanB	CPU Fan2	CPU Fan2A	Mem2R FanA	Mem2R FanB

- For Dual rotor fans, pick the RPM values for the higher RPM fan
- If there are multiple fans in same zone, use average RPM

Determine Model Coefficients



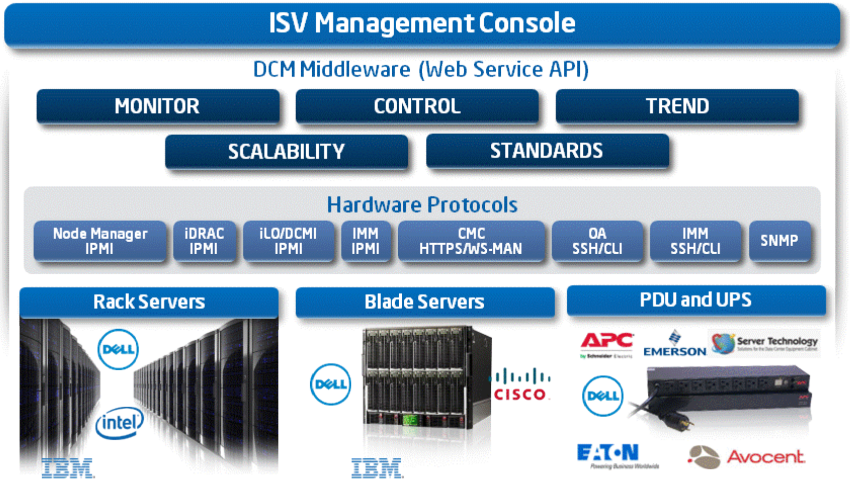
- Calculate the coefficients using the Intel® Management Engine (Intel® ME) SMBus Diagnostic Console (MESDC)
- Calculated coefficients may be directly entered into the appropriate Intel ME configuration file using the FITC tool

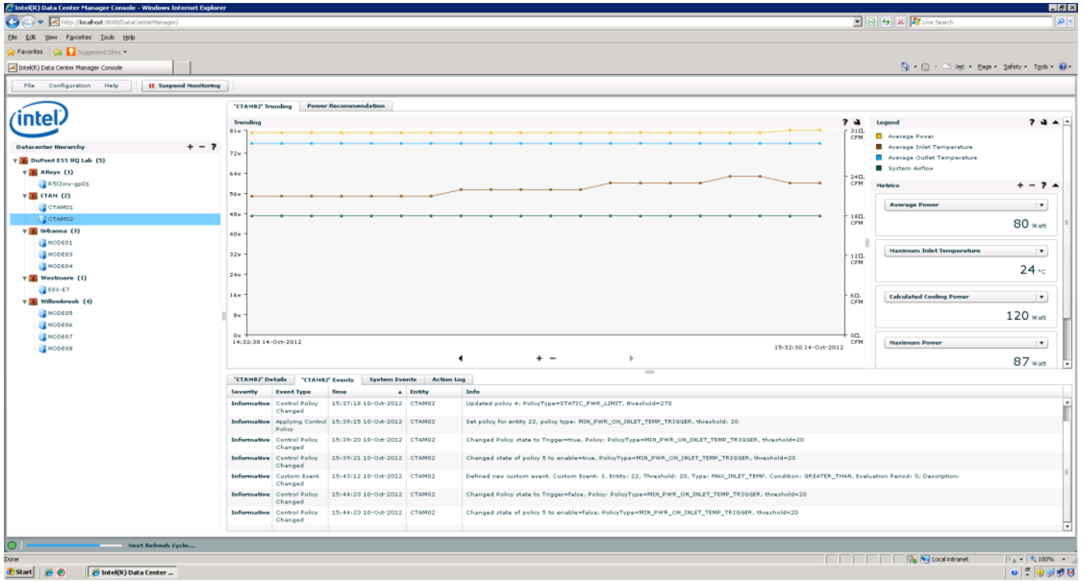
Power Thermal Usage Models and Use Case

- Monitoring & Reporting
- Thermal Modeling and Predictive Analysis
- Dynamic Controls

Reporting and monitoring – Use Intel Data Center Management Software

- Real time monitoring, reporting and set alerts and policies at server/rack/room level
- Aggregated control and trending
- Metrics monitored:
  - Inlet temperature
  - Outlet temperature
  - Power
  - Airflow





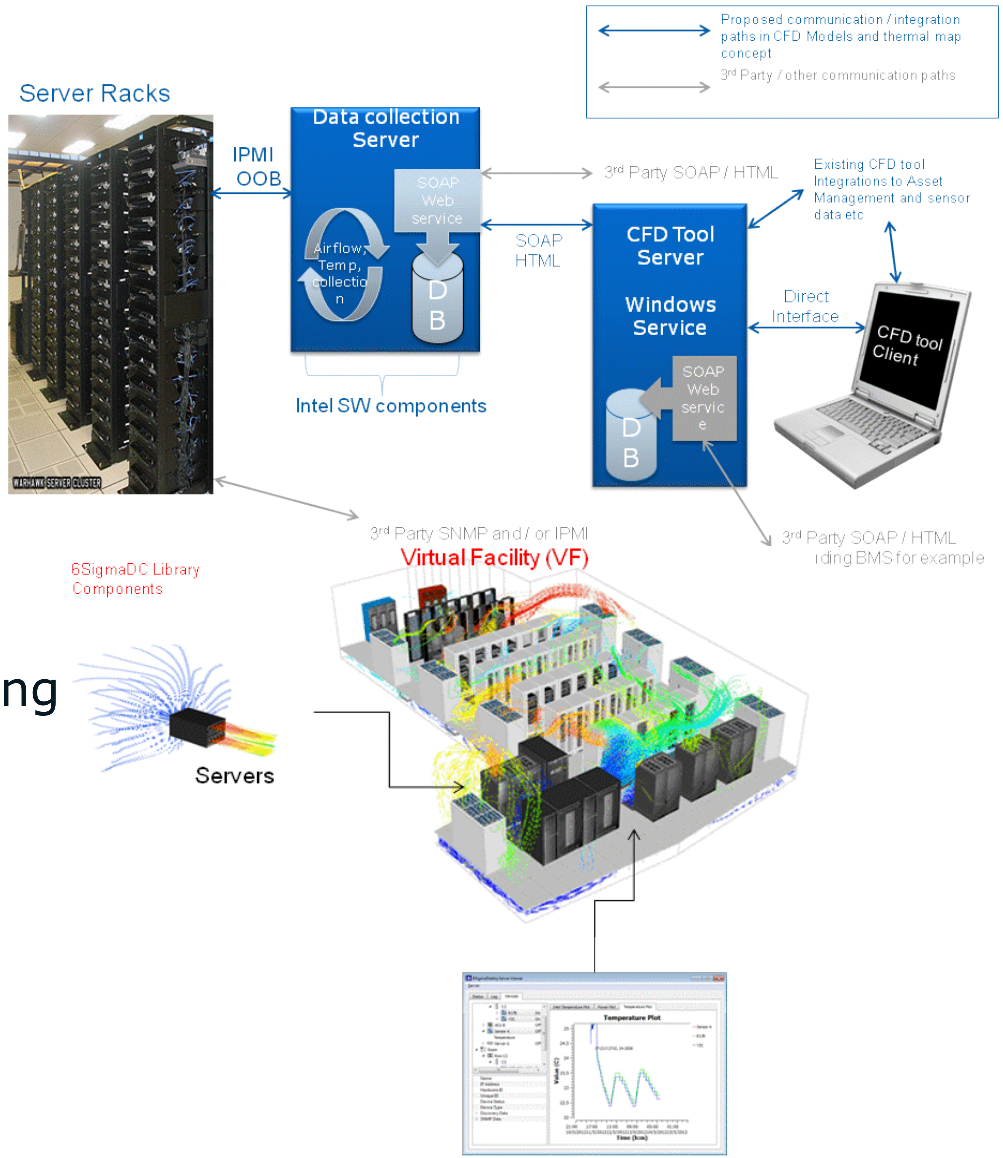
If you can't visualize it – you can't manage it!

Group Level Energy Management using DCM

Monitor	• DC thermal profile • Server health monitor (component temp, safe margin) • Dashboard and trending • Anomaly detection
Analysis	• DC Inefficiency (hot spot, uniform inlet/outlet temperature, by-pass, re-circulation. • System throttling impact under HTA environment
Alert	• Platform thermal errors Alert • DC inefficiency alert (hot spot, overall cooling, by-pass, • re-circulation)
Mitigation	• Power policy to put server into Emergence Mode • Power policy to put server into SHUTDOWN mode.
Optimization	• Airflow re-balancing • Optimize CRAC CFM/supply temp set point • Minimize compute energy

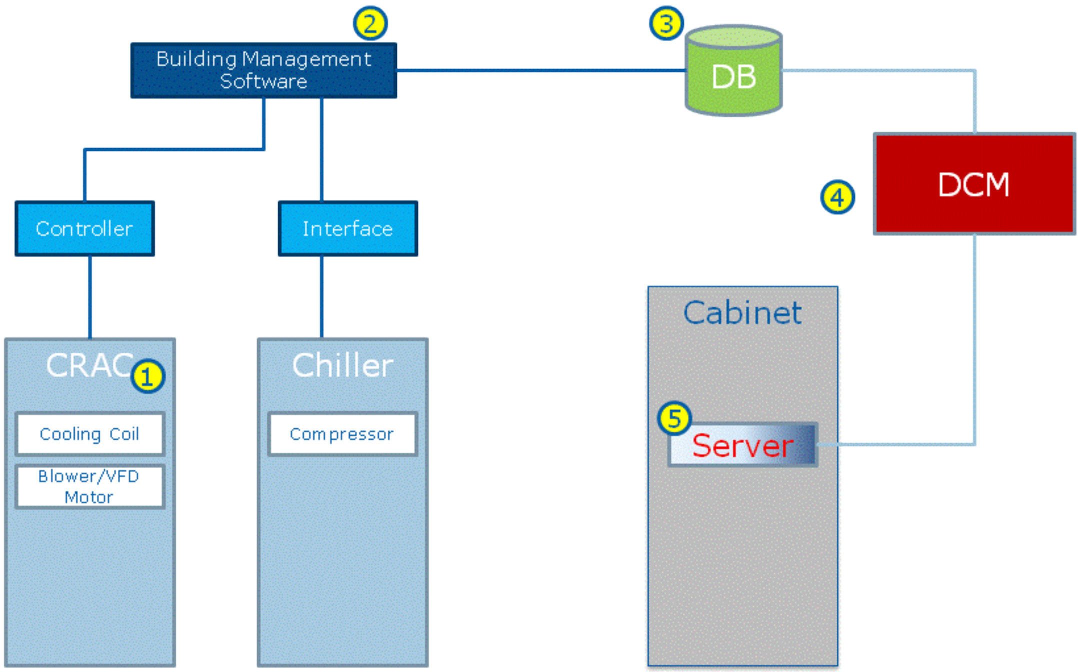
Thermal Modeling/Predictive Analysis

- Thermal and airflow maps are generated from the telemetry
- Models created used to
  - Identify problems
  - Troubleshooting
  - Management
- Models created can be tested before they are applied in the real world
- Simulate the impact to DC cooling when server based on future platforms are integrated into the DC



Dynamic Control / Cooling Balance

- Provide temperature and airflow information to Building Management Software (BMS)
- Use the temperature readings to modulate cooling fluid control valve
- Use the airflow readings to control fan speed of the CRAC



Demonstrate Potential Savings

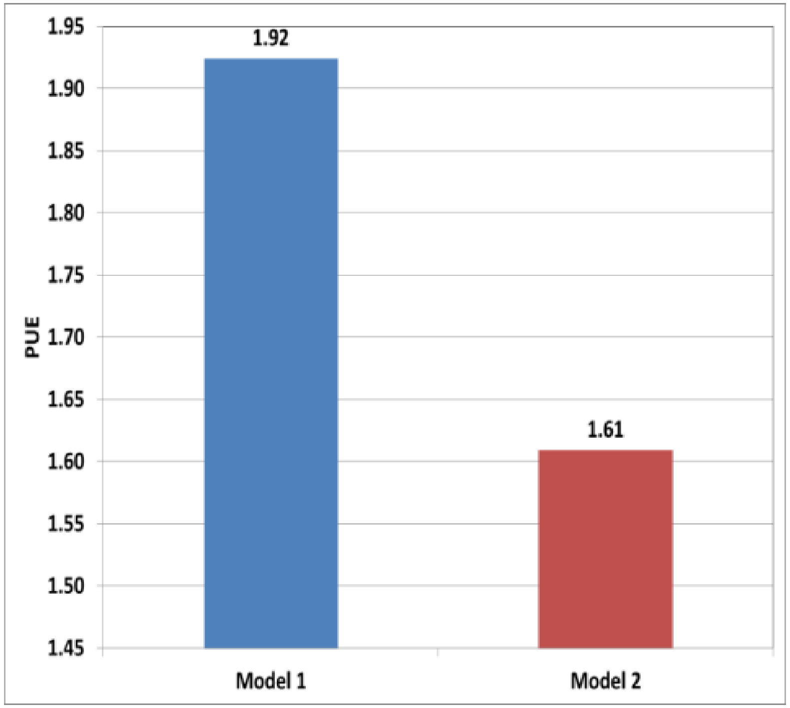
- Romonet* Software Suite is used to simulate the facility energy consumption
- Improvement in cooling efficiency improvement achieved by:
  - Airflow:
    - Matching airflow supply to airflow consumed by IT equipment
    - Fan Laws: Power ~ $Flow^3$
  - Temperature
    - Moving the cooling control from return side air temperature to supply air temperature.
    - Warmer operating temperatures allow significantly lower energy use in chiller
    - Lower airflow allows higher room Delta Ts; again better chiller efficiency

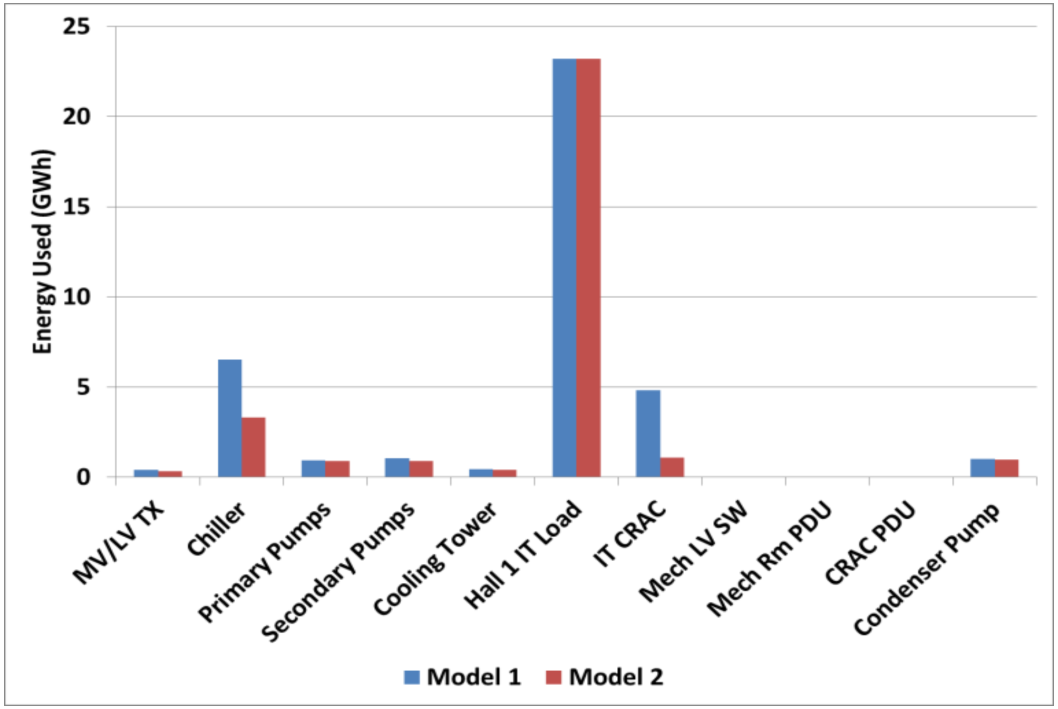
Case Study Scenarios

Data Center Scenario	Characteristics
Model 1	• Cooling Design: Chilled Water, Cooling Tower • Airflow Management: Hot/Cold aisle layout, No Containment • Fixed speed fans in CRAC • 21°C Return air temperature control
Model 2	• Cooling design: Chilled Water, Cooling Tower • Airflow Management: Hot/Cold aisle layout, Containment • Variable speed drives fans in CRAC (Match server airflow demand to airflow supplied by CRACs) • 21°C Supply air temperature control

- A hypothetical 1-MW facility is used to demonstrate the potential energy savings

Energy Consumption Comparison

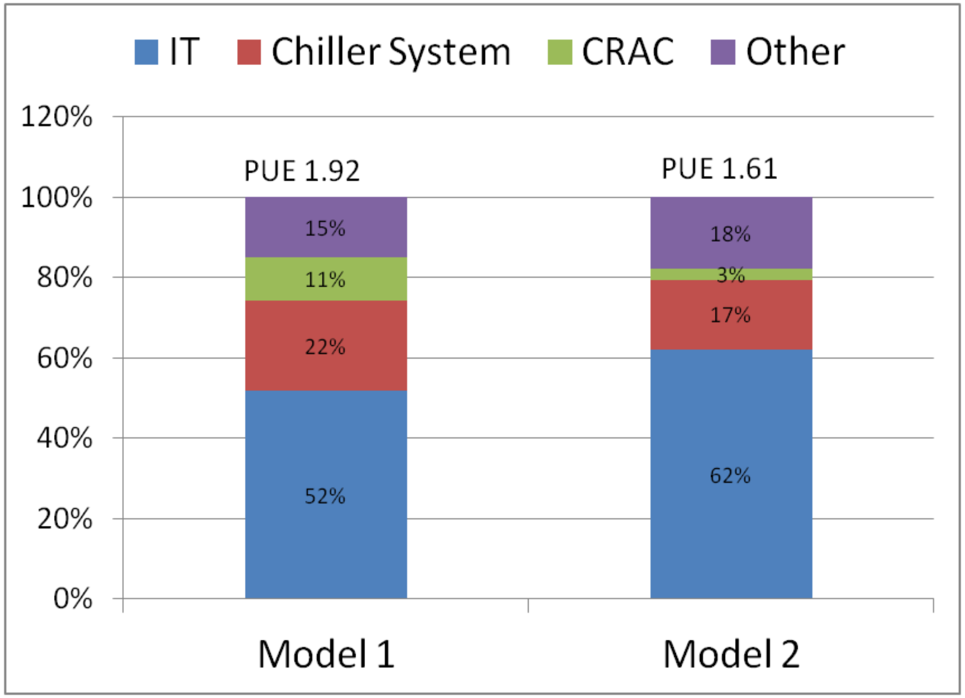




- Energy consumption of chilled water system is reduced by 35% (that is, 10 GWh to 6.5 GWh) by containment and moving temperature control from return side to supply side
- Energy consumption of CRAC is reduced by 77% (that is, 4.8 GWh to 1.1 GWh) moving to Model 2 that uses variable speed drive fans in the CRACs and by matching airflow demand of the servers to the air supplied by the CRACs
- Annual PUE reduced from 1.92 to 1.61

* Hypothetical 1-MW facility is used to demonstrate the potential energy savings possible by using Romonet* Software Suite

Overall Energy Consumption



- Chilled water system as a percentage of DC energy consumption reduces from 22% to 17% for model 2
- Overall energy consumption for the CRACs as a percentage of DC energy consumption reduces from 11% to 3% for model 2
- With a fixed power budget, less power consumed for cooling results in a higher power budget available for the IT equipment

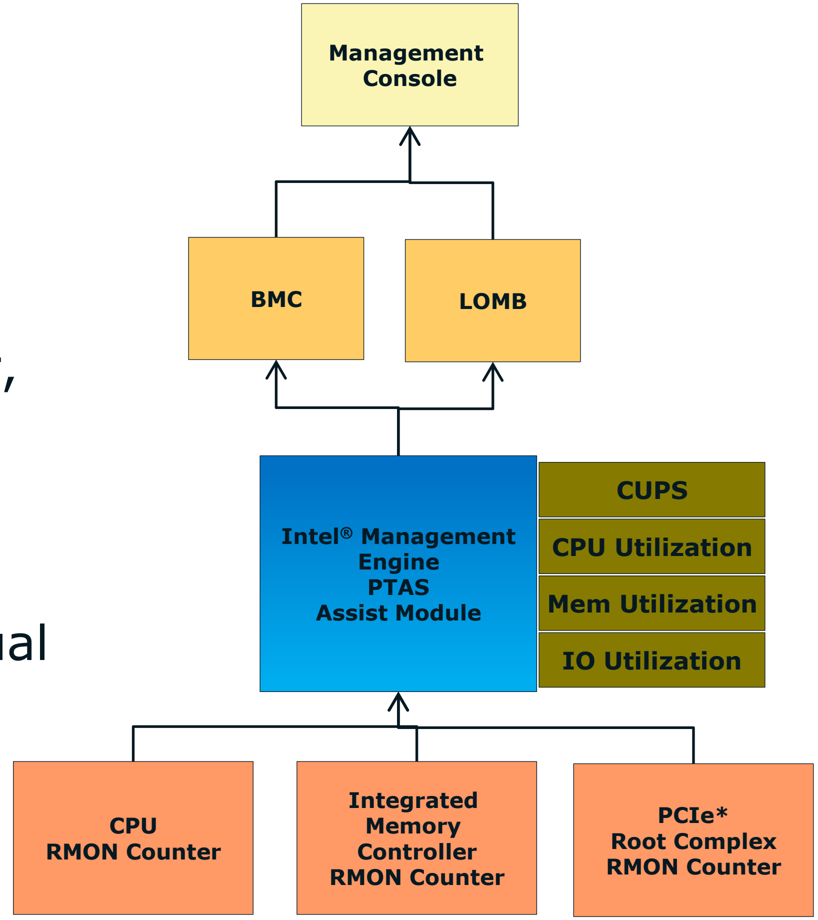
* Hypothetical 1-MW facility is used to demonstrate the potential energy savings possible by using the technology.

Agenda

- Problem Statement
- Methodology for Addressing Cooling in-efficiencies with PTAS
- Platform Enabling details
  - PTAS Thermal
    - Enabling steps
    - PTAS Thermal Usage
    - Potential energy savings
  - PTAS CUPs
    - Validation steps
    - PTAS compute usage
- Summary and Call to Action

PTAS – CUPS (Compute Usage per Second)

- Set of RMON counters that enable the tracking and reporting of compute utilization of the platform.
- Reduces the need for expensive OS based software tools
- Reduces cost of data integration – Enables real time monitoring of power, thermal and compute
- Introduction of a universal composite utilization metric
- User has the option of providing manual Load Factors or utilizing automatically determined values.
- No enabling required



PTAS Workload (CUPS) Overview

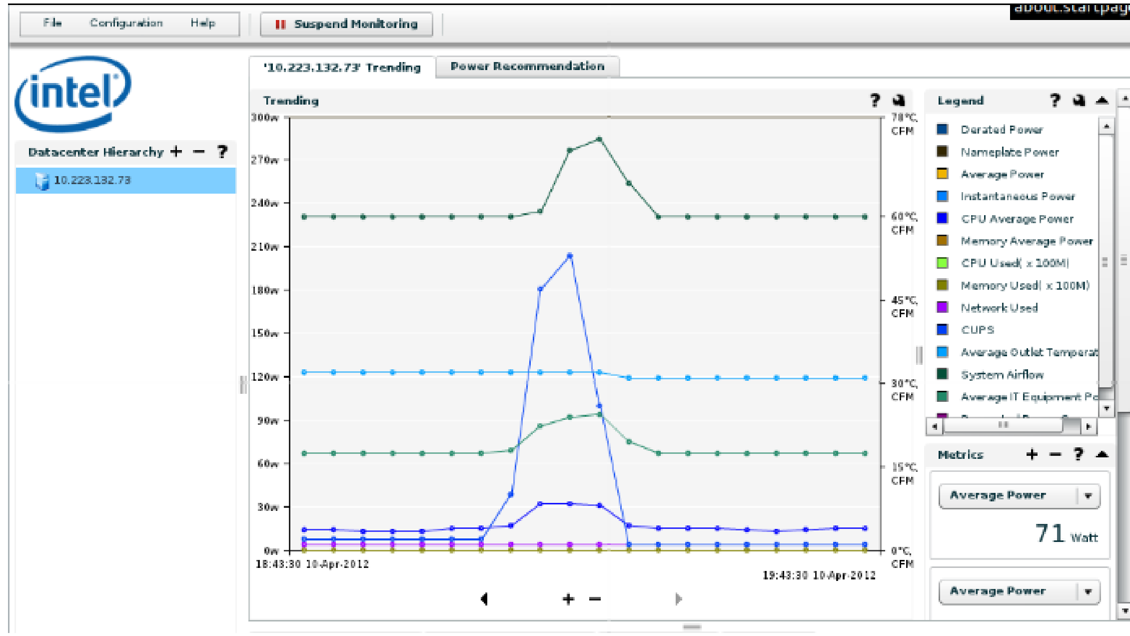
- The Intel® Management Engine will compute a CUPS Index based on the individual CPU, Memory, and I/O CUPS Indexes, adjusted by a Load Factor

$$CUPS = L_c W_c + L_m W_m + L_i W_i$$

- Workloads that run in data centers exhibit behavioral trends in favor of specific resources:
- Load factors adjust for these workload trends
- These can be adjusted manually (via commands) or automatically with FITc configured defaults

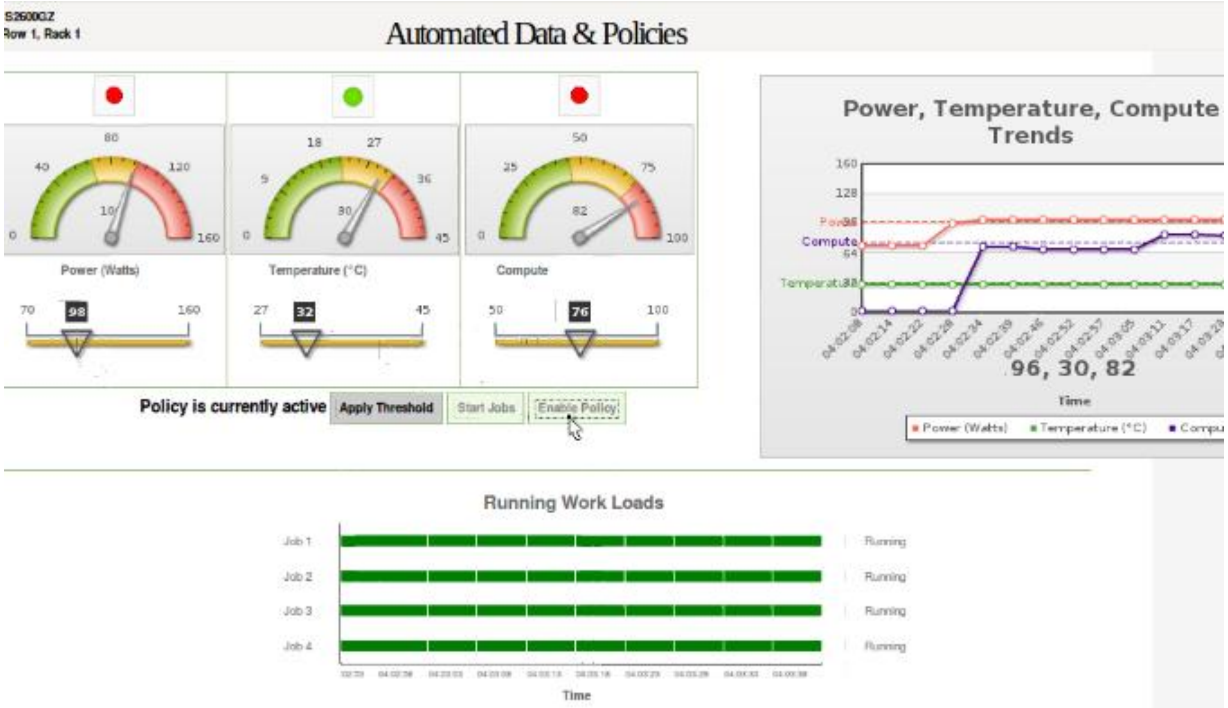
Usage - Reporting and Monitoring

- Centralized and Real time data monitoring, reporting and set alerts and policies at server/rack/room level
- Aggregated control and trending
- Metrics monitored:
  - CPU utilization
  - CUPS
  - Memory utilization
  - I/O utilization



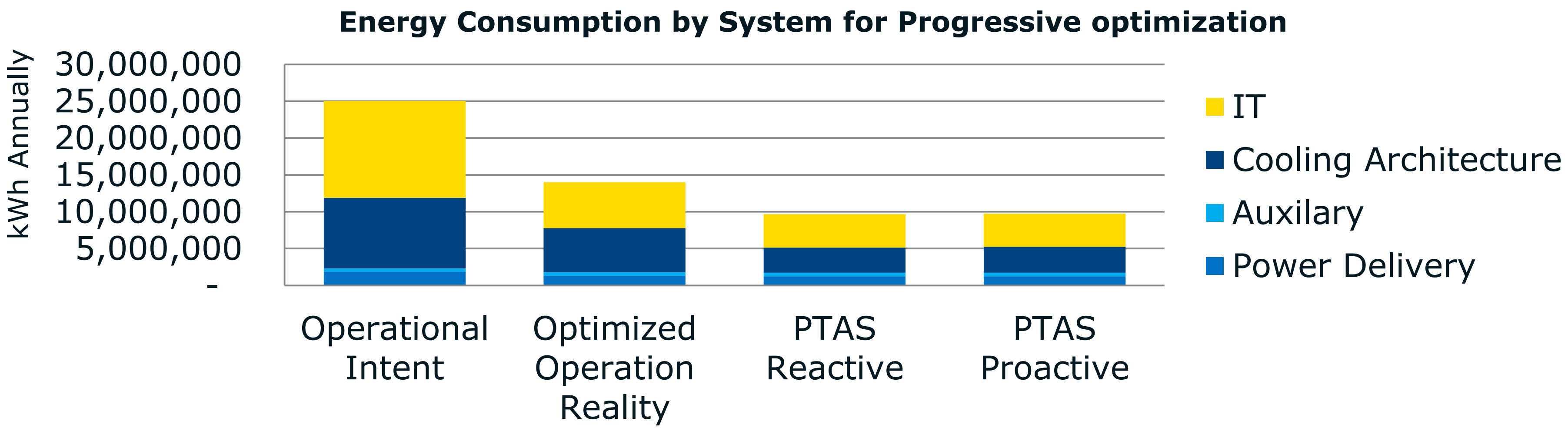
Usage – Reporting, Monitoring & Alerts

- Power thermal and compute data monitoring, dashboard and trending
- Anomaly detection and alarms
- Proactive alerts based on thresholds



Usage - Analytics

- Model w/ 4 configurations – Base, Optimized, PTAS reactive, PTS proactive
- Apply server refresh to bring Base to Optimized.
- Apply PTAS analytics to get increased energy efficiency and availability



Operational Intent	Optimized Ops	PTAS Reactive	PTAS Proactive
• Base configuration • 7000 sq. ft • Hot and Cold Aisle Layout • 300 cabinets/6300 servers • 14 ACU/CRACs • PUE 1.5-3.0	• Systems refreshed and workloads consolidated for compute	• PTAS policies to balance workload using uniform outlet profile	• Minimize compute energy to place system in sleep state • Feedback to CRAC for communication & control • Reduce # of CRACs

* Model built internally w/ 3rd party DC predictive modeling tool (Romonet* SW suite)

Agenda

- Problem Statement
- Methodology for Addressing Cooling in-efficiencies with PTAS
- Platform Enabling details
- Summary and Call to Action

Summary

- Mismatch between airflow supplied by the CRAC fans and airflow required by the IT equipment is a source of cooling inefficiency in a Data Center.
- Server inlet temperature and server airflow as two thermal control points available from the server to DCM tool via IPMI commands.
- The integrated server airflow can be used to determine CRAC fan speeds, thereby improving cooling efficiency.
- This, when combined with hot and cold aisle isolation, can eliminate bypass of cold air to hot aisle and re-circulation of hot air back to cold aisle,
- Use compute metrics to balance workloads to achieve a uniform thermal profile of the racks.

Call To Action

- Build PTAS enabled platforms
  - Start platform enabling activities
    - Increase platform value proposition
    - Get higher ASP systems !
- Enable Dynamic Scaling solutions
  - Use Data Center Manager (DCM) to start real-time monitoring and management
- Be energy efficient !

Abbreviations and Acronyms

DCM	Intel Data Center Manager Software
CRAC	Computer room air-conditioning unit
PTAS	Power Thermal Aware Solution
HVAC	heating, ventilation, and air conditioning
UPS	Uninterruptible power source
PUE	Power usage effectiveness
CUPs	Compute usage per second
IPMI	Intelligent Platform Management Interface
ACU	Air Conditioning Unit
BMS	Building Management Software

Backup

PTAS Usages & Benefits

Usage Model	Use Cases	Benefits
Real-time Monitoring & Reporting	• Centralized & Real time data monitoring • Power, thermal & compute - dashboard & trending • Inventory/Asset Location • Anomaly detection & alarms	• Real-time Monitoring & Visibility (you can manage what you can't measure) • DC Thermal & Power mapping • Typing power requirement and heat generation to work performed to make intelligent decisions
Analytics & Alerts	• Proactive alerts based on thresh-hold • Reactive/Proactive failure Analysis	• Problem analysis, handling & prevention • Workload, cooling & power issues identification (recirculation, bypass) • Workload placement recommendation
Control: cooling Optimization	• Manage thermal events • Platform power management • Capacity/Deployment planning • CFD modeling • Supply side optimization	• Increase cooling efficiency (2.4% fan power, 36% CWS) • Reduce wasted energy inefficiency with bypass, recirculation, ACU oscillation (90%) & crosstalk (20-40%) • Reduce DC safety margins (7c-over build) • 2 way ACU communication & control
Control: compute Optimization	• Workload placement and relocation on events • Server ranking • Uniform Output Profile (UOP) • Minimize Compute Energy (MCE)	• Lower cooling Opex (8-16%) • Lower cooling Capex (25%) • Reduce/eliminate external sensor instrumentation ($1000/svr or sqft) • PUE reduction • Minimize stranded capacity • Workload characterization / visibility • Workload placement optimization

Increase DC energy efficiency by 30-40%

PTAS Schedule & Collateral

- Schedule:
  - PTAS is scheduled to be released in parts:
    - PTAS Thermals will be available at the Alpha release of ME FW
    - PTAS CUPS will be available at the Beta release of the ME FW
  - Alpha is currently trending to the mid-May timeframe
  - Beta is currently trending to the August-September timeframe

- Collateral:
  - Currently there is some information on the PTAS feature in the existing Grantley collateral:
    - Intel® Intelligent Power Node Manager 3.0 External Interface Specification using IPMI – #513973
    - Intel® Server Platform Services 3.0 E5 Firmware External Product Specification – #516470
    - Intel® Server Platform Services 3.0 E5 – Extended Services Integration Guide – #520427
  - Some of the existing information may be incomplete or pending but this is being currently updated. Expect future releases of the collateral to have more complete information on the feature