GPU and AI Infrastructure Management with AMI® Data Center Manager

Whitepaper Commissioned by AMI

February 2025

Author

Rami Radi: Sr. Product Manager and Solution Architect

Table of Contents

Executive Summary

As data centers grow in size, complexity, and power demands, managing energy efficiency, AI infrastructure, firmware tracking, and infrastructure costs has become a major challenge. Traditional Data Center Infrastructure Management (DCIM) solutions typically focus on facility-wide monitoring, while server and workload management tools are limited to optimizing CPU, GPU, and memory resources. However, neither offers the device-level insights necessary to optimize power consumption, cooling efficiency, firmware consistency, and component health.

AMI® Data Center Manager (DCM) fills this gap by offering real-time power monitoring, AI and GPU infrastructure insights, automated firmware tracking, and intelligent device-level monitoring. Designed to complement—not replace—DCIM and server management solutions, AMI DCM helps data centers reduce costs, improve sustainability, and maximize infrastructure utilization.

This whitepaper explores the major challenges facing data centers today as well as some of the key features and use cases of AMI DCM, demonstrating its potential significant financial and operational benefits.

The Challenges Facing Modern Data Centers

Modern data centers face several growing challenges that impact scalability, efficiency, and cost-effectiveness. These include:

1. Scaling on Demand

Modern AI and HPC infrastructure devices are both powerful and costly, requiring granular monitoring. Even the failure of just one or two nodes can cause catastrophic disruptions to workloads, emphasizing the critical need for device-level management in these environments. The ability to monitor and optimize power and thermal footprints at the device level is critical for maintaining cost-effective growth.

2. Hybrid Environments and Multi-Vendor Complexity

Many organizations work within hybrid cloud environments, integrating on-premises, private, and public cloud infrastructures, while managing complex devices that demand close monitoring. This introduces additional complexity, as data centers rely on servers, networking, storage, and cooling systems from multiple vendors. Managing such diverse environments without unified device-level visibility leads to inefficiencies.

3. Diverse Silicon Architectures and AI Compute Demands

Supporting Intel, AMD, and ARM-based servers (including Ampere and NVIDIA Grace) adds complexity to infrastructure management. Moreover, high-density GPU clusters such as NVIDIA’s GB200 NVL72 require precise power, health, and thermal monitoring and management to prevent power waste and hardware degradation.

4. Preventing Downtime and Ensuring System Health

Unexpected hardware failures are major causes of unplanned downtime in data centers. A lack of proactive health monitoring often leads to reactive fixes, increasing operational costs and service disruptions. Without device-level monitoring of component health, failure patterns, and early warning signs, organizations are left responding to outages instead of preventing them.

Minimizing downtime is not just about redundancy, it’s about identifying risks early. By monitoring power anomalies, thermal hotspots, and hardware degradation trends, data centers can take preventive action before failures occur, significantly reducing operational disruptions and revenue loss.

5. Energy Efficiency and Sustainability Goals

With rising energy costs and increased sustainability mandates, data centers must optimize PUE and reduce power waste. Many facilities, however, lack real-time, device-level power and cooling insights, making it difficult to manage sustainability efforts effectively. By optimizing cooling systems, significant energy savings can be achieved. For instance, adopting liquid cooling solutions can lead to lower overall energy consumption and reduced environmental impact (Sources: Deloitte, 2024; Boyd Corporation, 2024).

6. Firmware Tracking and Management

Data centers often run a mix of firmware versions across thousands of devices, leading to inconsistencies that can impact reliability, performance, security, and compliance. Identifying and automating firmware consistency checks across multi-vendor environments is crucial for operational efficiency. Moreover, having a reliable means to manage and update device firmware in large batches is a necessity that is growing more complicated by the day.

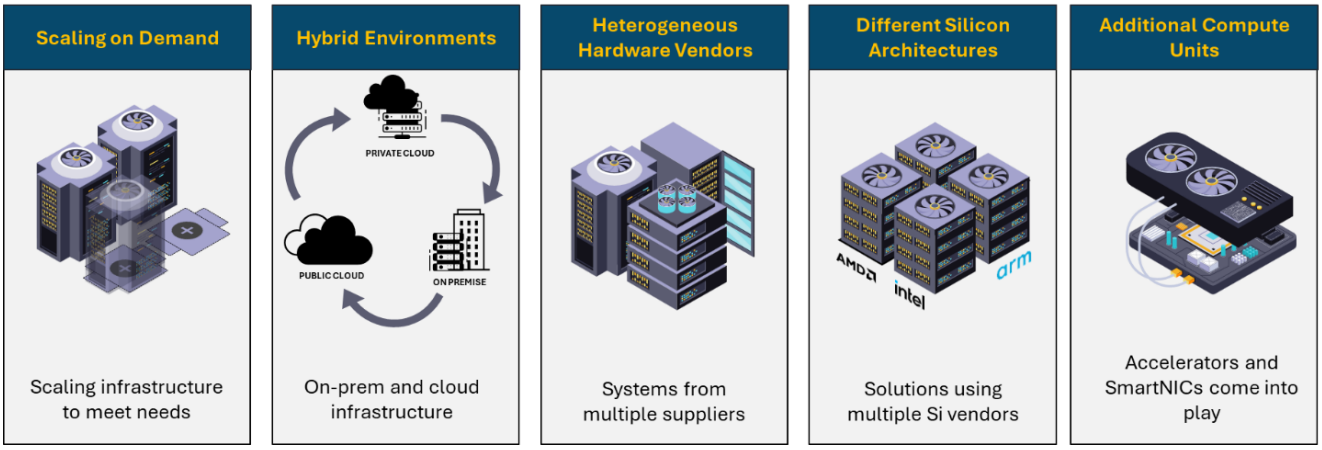


Figure 1: Top data center challenges that AMI DCM addresses.

Overview of AMI Data Center Manager

AMI DCM manages and monitors each device at a granular level while providing a comprehensive view of the entire data center, offering actionable insights and improvements across both levels. As an out-of-band solution, it does not require any intrusive changes to existing infrastructure. Unlike in-band monitoring tools that require an agent to run on each server—potentially impacting performance and adding operational complexity—AMI DCM communicates directly with devices using standard protocols such as IPMI, Redfish, SNMP, and SSH. It supports a wide range of devices, including servers, networking, storage, power, and cooling equipment, across various architectures and vendors.

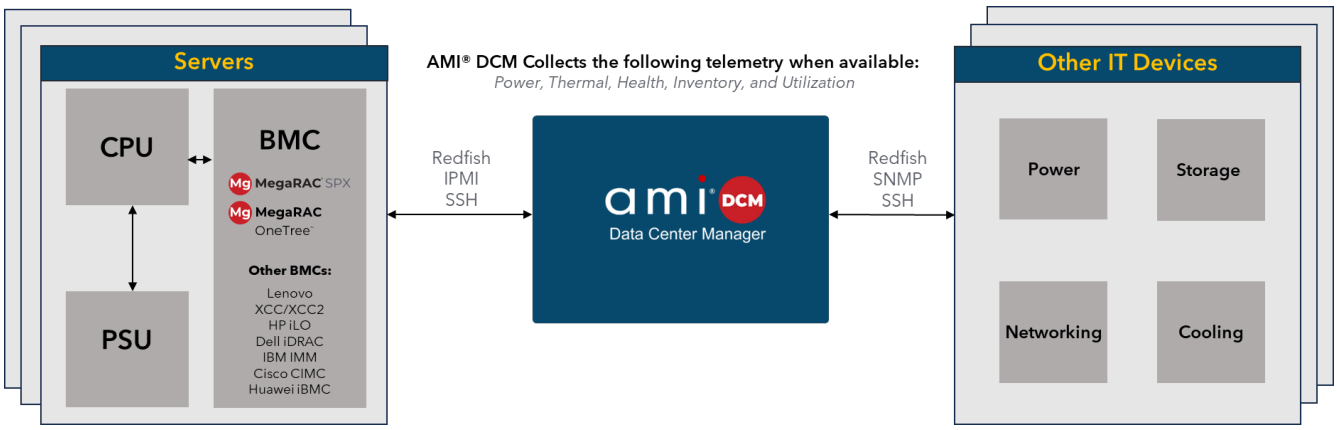


Figure 2: How AMI DCM communicates with a broad range of devices.

Furthermore, AMI DCM is a mature, field-proven solution that has been developed in collaboration with major technology partners for years, ensuring support for each new generation of data center devices. Its ongoing evolution continues to provide organizations with comprehensive, scalable, and vendor-agnostic data center management capabilities.

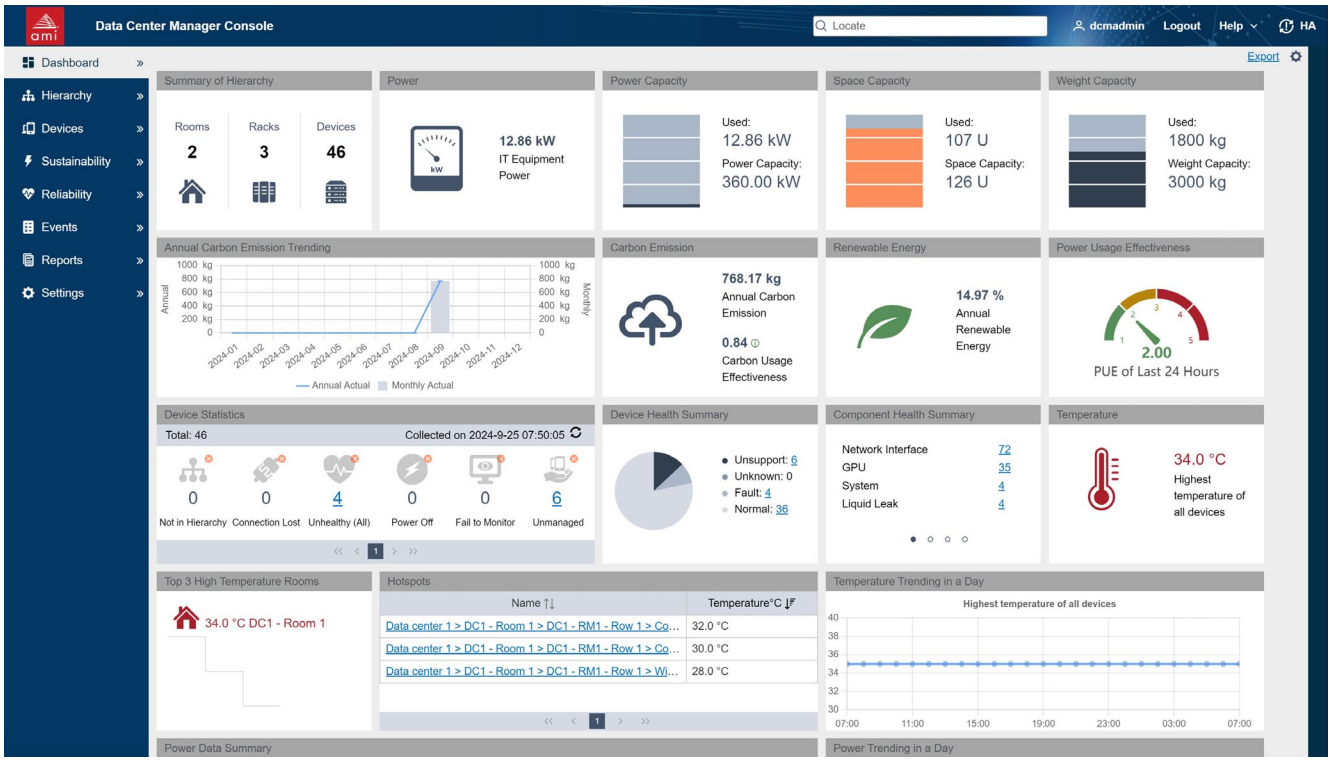


Figure 3: AMI DCM Web console dashboard showing a holistic overview of a data center.

Key Features of AMI Data Center Manager

1. Energy and Cooling Optimization

AMI DCM provides real-time power and thermal monitoring, eliminating the need for smart PDUs, and enabling optimized energy use at device, rack, and facility levels. By continuously collecting and analyzing power draw and thermal data, AMI DCM enables operators to proactively adjust cooling strategies, balance power distribution, and minimize energy waste. Additionally, AMI DCM calculates and tracks key sustainability metrics such as Power Usage Effectiveness (PUE) and Carbon Usage Effectiveness (CUE), ensuring data centers meet efficiency goals and regulatory requirements. These insights help data centers reduce operational costs, minimize carbon footprints, and extend the lifespan of IT infrastructure by preventing overheating and energy inefficiencies.

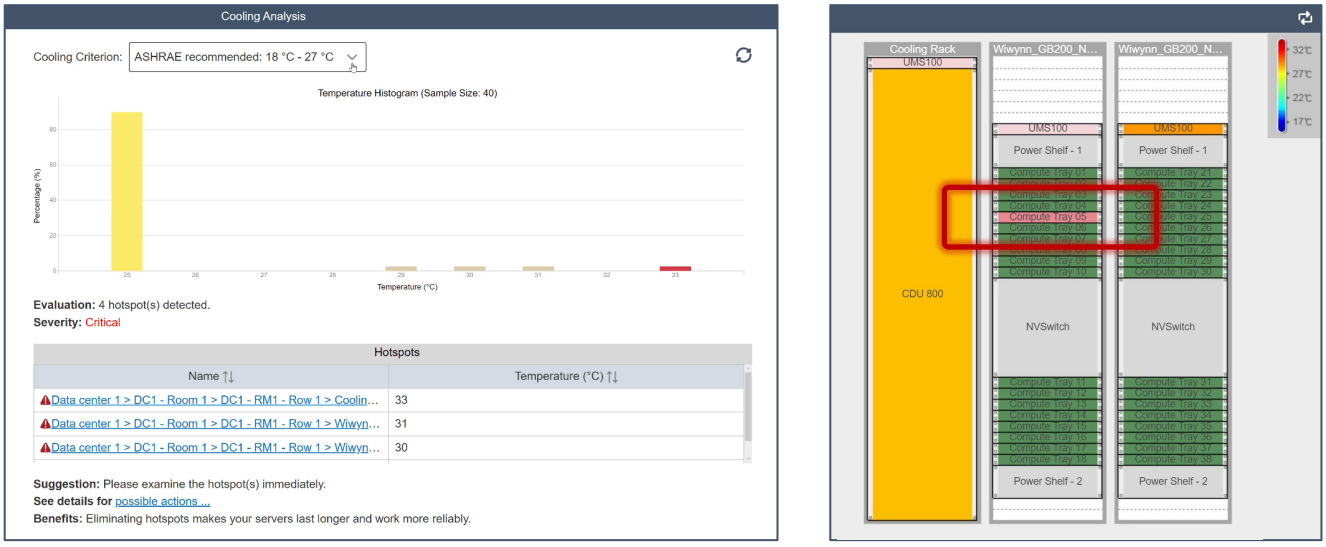


Figure 4: Identifying opportunities to safely increase set point temperature (left). Identifying thermal hotspots due to insufficient cooling or health issues (right).

2. AI and GPU Infrastructure Insights

With the increasing adoption of AI and high-performance computing (HPC) workloads, GPUs have become one of the most energy-intensive and thermally demanding components in data centers. AMI DCM now provides granular monitoring and management of NVIDIA data center GPUs, including the GB200 NVL72 platform, allowing data center operators to track power consumption, thermal behavior, and utilization trends in real-time.

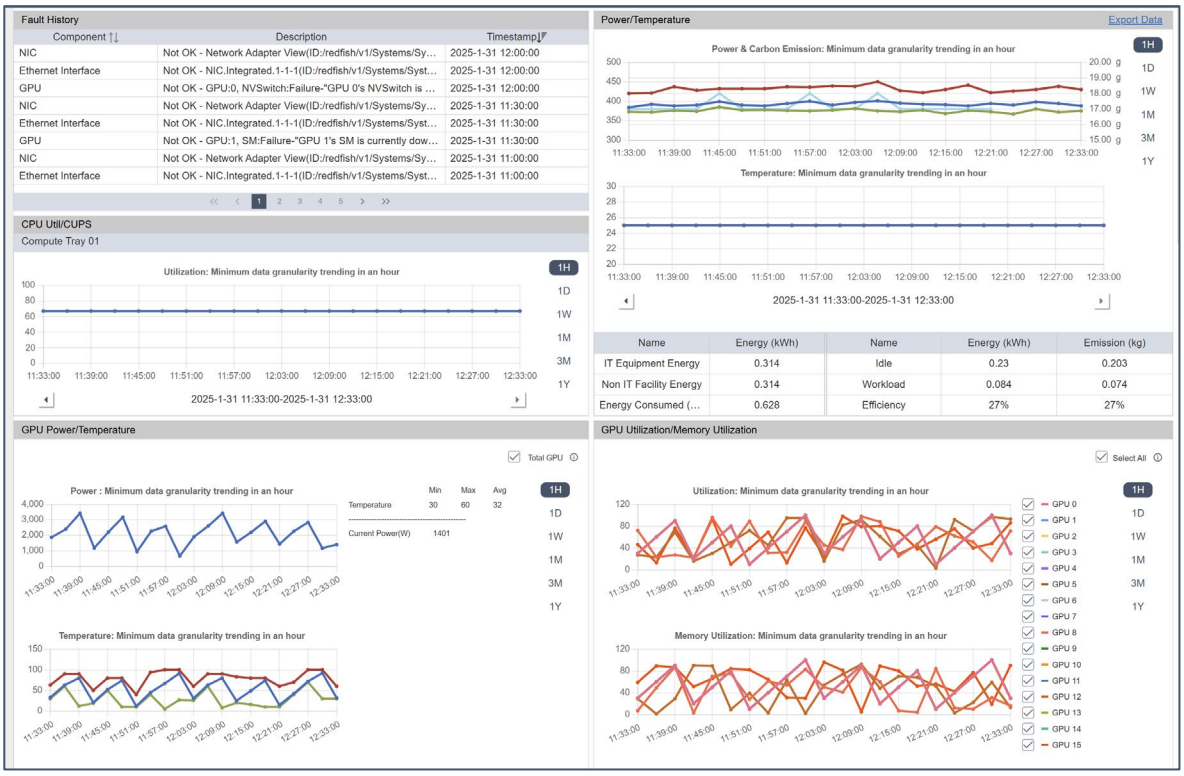


Figure 5: Server view showing server, CPU and GPU utilization, power, thermals, and carbon emissions.

By leveraging these insights, operators can prevent thermal throttling, optimize power allocation, and improve cooling strategies—all while ensuring AI and HPC workloads run at peak performance without unnecessary power waste. AMI DCM also supports liquid cooling environments, helping data centers integrate and manage next-generation cooling technologies to enhance efficiency and reliability in high-density AI clusters.

3. Firmware Monitoring and Consistency Management

Managing firmware across a heterogeneous data center is a significant challenge, as different devices from multiple vendors may run on varying firmware versions. AMI DCM provides firmware tracking, enabling IT teams to identify discrepancies in firmware versions across servers, storage devices, power equipment, and cooling systems. Inconsistent firmware versions can lead to hardware instability, degraded performance, and operational inefficiencies. By detecting and reporting firmware mismatches and allowing data center operators to proactively schedule Redfish-based server updates in batches, can reduce unexpected issues caused by outdated or mismatched firmware.

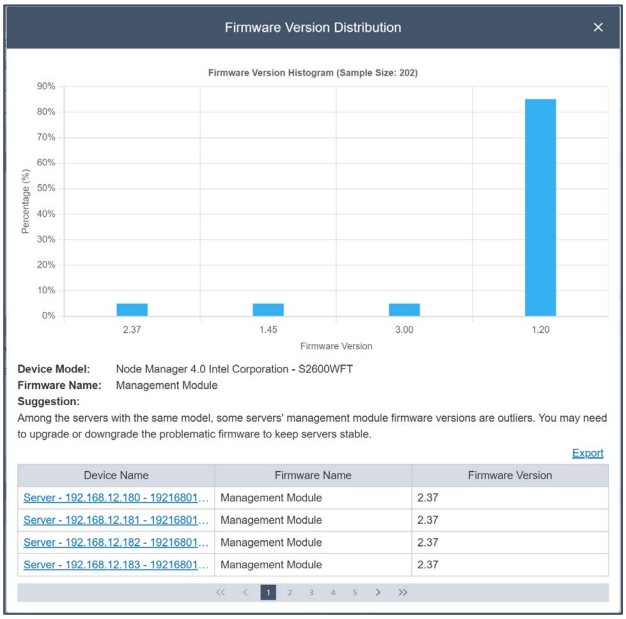


Figure 6: Server management module firmware version distribution for servers of the same model.

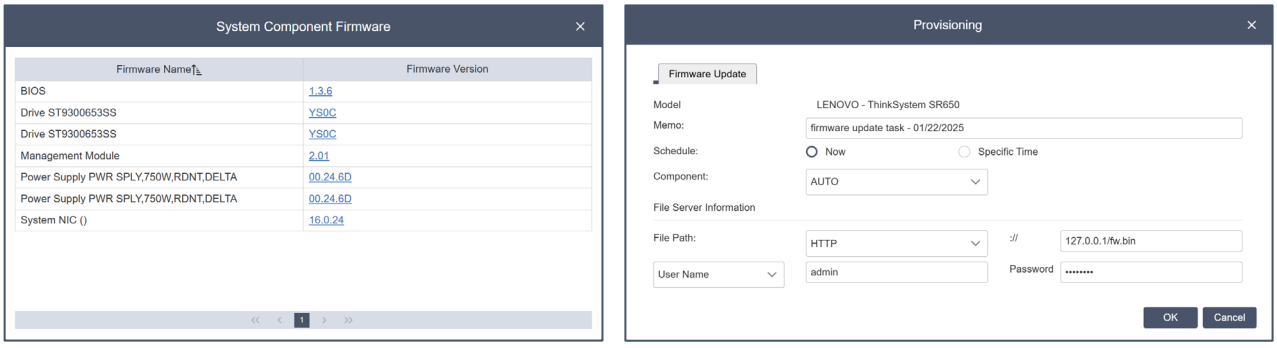


Figure 7: Firmware versions of the different components of a server, and server firmware updating.

4. Multi-Vendor, Multi-Architecture Compatibility

Unlike vendor-specific tools, AMI DCM is hardware-agnostic, and provides standardized monitoring and management across diverse IT environments. It supports Intel, AMD, and ARM-based server architectures, and supports a wide range of networking, storage, power, and cooling devices. With native support for industry-standard protocols such as IPMI, Redfish, SNMP, and SSH, AMI DCM eliminates fragmented monitoring and providing a unified view of all data center IT devices. This ensures that organizations can centrally manage their infrastructure, without being locked into a single vendor's ecosystem, or without spending on multiple vendor solutions.

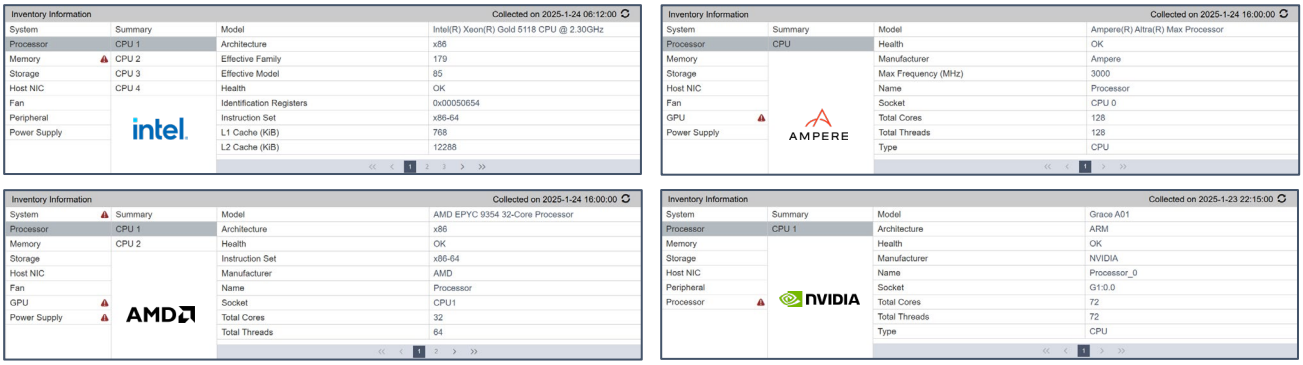


Figure 8: Server inventory view showing support for Intel, AMD, Ampere, and NVIDIA CPU architectures.

5. Infrastructure Cost Optimization

Data centers frequently struggle with overprovisioned, underutilized, or inefficiently distributed resources. AMI DCM helps optimize infrastructure by identifying and addressing inefficiencies, including low-utilization servers, inefficient cooling practices, and unnecessary Smart PDU deployments.

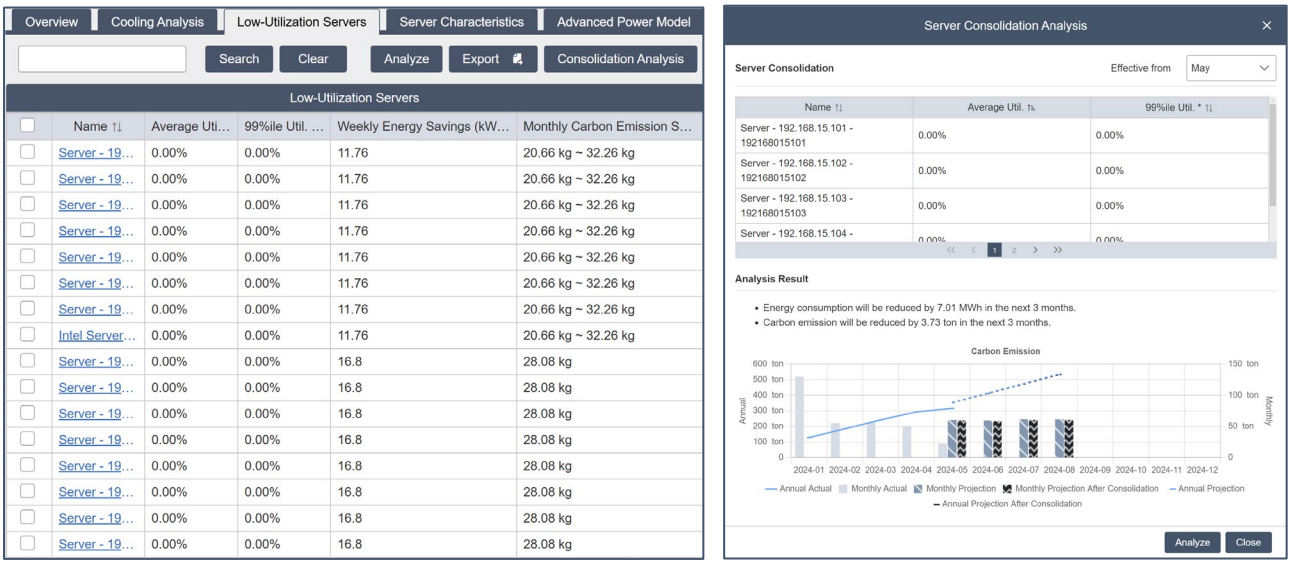


Figure 9: Identifying underutilized servers (left). Potential energy and emission gains from consolidation (right).

By providing accurate power and utilization, AMI DCM helps operators determine which servers can be decommissioned or consolidated, how rack density can be improved, and where energy savings can be achieved. Additionally, by replacing expensive Smart PDUs with software-based power monitoring, data centers can significantly reduce capital expenses while maintaining visibility into power consumption. These optimizations help data centers avoid unnecessary expansions, reduce overhead costs, and extend the lifecycle of existing infrastructure, improving long-term ROI.

6. Multiple Integration Options

AMI DCM's capabilities are accessible through multiple interfaces, including a Web Console, RESTful APIs, Apache Kafka data streaming, and iFrame integration. This allows organizations to integrate AMI DCM's insights into their existing solutions, ensuring maximum operational value.

These integration options have enabled AMI DCM to be integrated into multiple commercial solutions, including DCIM, server management, and other enterprise solutions, further expanding the ecosystem and strengthening interoperability across modern data centers.

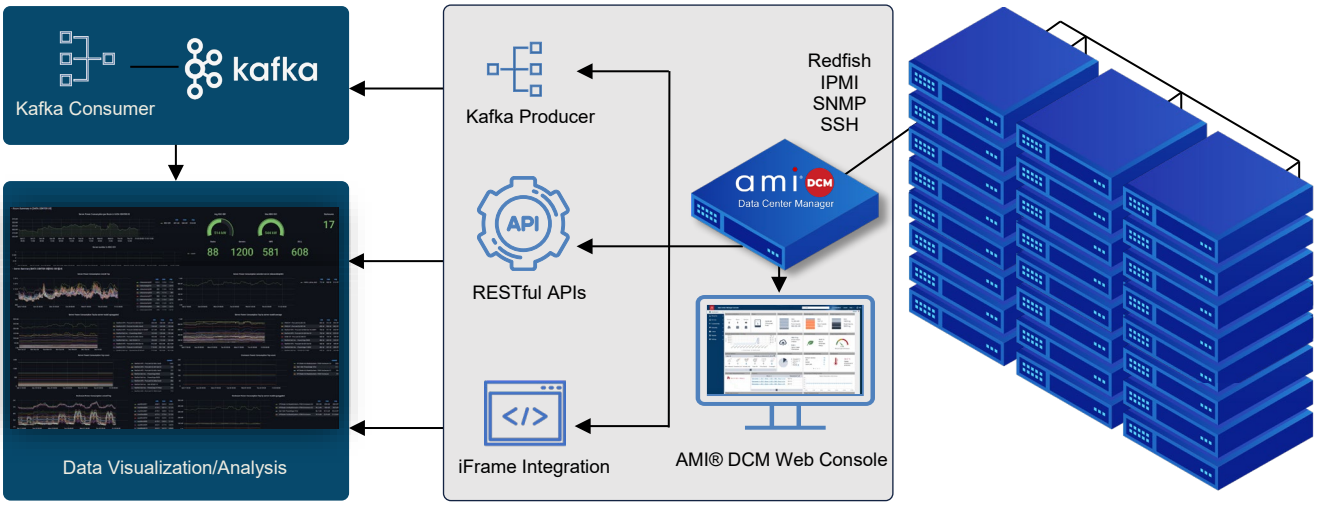


Figure 10: AMI DCM interfaces and integration points.

Use Cases and ROI Analysis

To calculate the ROI of using AMI DCM, let us look at a medium-sized data center containing 5,000 servers across 250 racks, operating with a PUE of 1.5, and with a power cost: $0.12 per kWh.

1. Identifying and Eliminating Underutilized Servers

Many data centers operate idle or underutilized servers ('zombie servers') that consume power, generate heat, and occupy rack space without contributing to operations. Studies such as those from NREL, estimate that 10–30% of servers in data centers are underutilized (Source: EPRI, 2024).

- AMI DCM identifies these servers in real time through:
- Power and CPU utilization tracking to flag inefficiencies.
- Long-term trend analysis to distinguish temporary vs. persistent low usage.
- Idle power consumption insights to quantify wasted energy.

Decommissioning or consolidating these servers leads to:

- Lower energy and cooling costs by reducing power consumption.
- Deferred infrastructure expansions by freeing up rack space.
- Minimized maintenance costs by removing unnecessary hardware.

ROI Calculation:

- *500 underutilized servers (10% of total).*
- *Power cost per idle server/year = $158*
- *Space, cooling, maintenance, and warranty savings per server = $450*
- *Total ROI: (Power cost per idle server + Space and cooling savings per server + Maintenance cost per server + Warranty cost per server) × (Number of servers that can be shutdown)*
- ***Total ROI: ($158 + $450) × 500 = $304,000 per year.***

2. Optimizing Cooling by Safely Raising Data Center Temperatures

Cooling systems account for a significant portion of energy consumption in modern data centers, with estimates suggesting that as much as 30-40% of the total energy is used for cooling (Source: EPRI, 2024). Yet, many data centers overcool equipment due to conservative temperature settings. AMI DCM enables smart cooling strategies by:

- Identifying safe thermal headroom in servers and racks.
- Preventing hot spots by analyzing airflow and rack temperature.
- Gradually raising setpoints while ensuring stability.

A carefully managed 3°C increase, driven by real-time thermal analysis, can reduce cooling costs by up to 10-15%, maintaining a consistent and conservative estimate.

ROI Calculation:

- Cooling cost reduction per 1°C increase: 6%
- Server power cost per year: $631
- Cooling percentage of total energy: 40%
- Total ROI per 3°C increase: (Cooling cost reduction per degree increase) x (# of servers) x (server power cost per year) x (assumed average PUE) x (percentage of total DC energy consumption for cooling) x number of degrees increased
- ***Total ROI per 3°C increase: (0.06 x 5000 x 631 x 1.5 x 0.4) x 3 = $340,740 per year***

3. Eliminating the Need for Smart PDUs

Smart PDUs provide rack-level power monitoring but cost $3,000–$4,000 per unit. Since AMI DCM offers device-level power monitoring, data centers can switch to cost-effective basic PDUs without losing visibility. With an average lifespan of approximately 10 years, replacing Smart PDUs with basic ones can lead to significant long-term savings. Key benefits include:

- Avoiding high CapEx costs of Smart PDUs while maintaining power insights.
- Reducing configuration complexity in large deployments.
- Consolidating power monitoring under a single-pane-of-glass with AMI DCM.

ROI Calculation:

- Basic PDU cost: $1,475/unit vs. Smart PDU cost: $3,275/unit
- 4 PDUs per rack (2 main, 2 redundant) across 250 racks
- Total savings over 10 years: (Smart PDU cost – Basic PDU cost) × (PDUs per rack) × (number of racks)
- ***Total ROI over 10 years: ($3,275 - $1,475) × 4 × 250 = $1,800,000 per year***
- ***Total ROI per year: $180,000***

4. Optimizing Rack Density to Reduce Expansion Costs

Data centers often resort to expanding their footprint by deploying additional racks and infrastructure, without optimizing their existing infrastructure. Instead, AMI DCM can help data centers optimize rack utilization, ensuring that existing resources are fully utilized before investing in new hardware. This has the following benefits:

- Defers costly expansions by maximizing rack efficiency.
- Reduces networking complexity and infrastructure sprawl.
- Improves cooling efficiency through better server placement.

ROI Calculation:

- Cost per rack (including space, networking, and PDUs): $15,000
- 10% increase in rack density avoids purchasing 25 new racks
- Total ROI = (Cost of a rack) × (Percentage of rack density increase) × (Total number of racks)

- ***Total ROI: $15,000 x 0.10 x 250 = $375,000 per year***

5. Reducing Business Downtime with Early Issue Detection

Downtime leads to revenue loss, SLA penalties, and service disruptions. AMI DCM reduces failures by detecting power anomalies, thermal issues, and hardware degradation early, allowing for proactive maintenance. If AMI DCM cuts downtime per failure by 1 hour (from 2 hours to 1 hour).

ROI Calculation:

- Annual probability of server failure: 5%
- Failure frequency per year: 4
- Downtime cost per hour: $1,000
- Total ROI: (Number of Servers) × (Probability of Hardware Failure) × (Reduction in Downtime) × (Cost of Downtime per Hour) × (Frequency of Failures per Year)
- ***Total ROI: 5,000 × 0.05 × 1 × $1,000 × 4 = $1,000,000 per year***

Summary and Conclusion

This whitepaper has explored the growing challenges in modern data center operations, highlighting the increasing complexity of scaling infrastructure, managing hybrid environments, handling multi-vendor hardware, optimizing AI and GPU resources, reducing power consumption, and ensuring firmware consistency.

AMI Data Center Manager (DCM) bridges the gaps between DCIM, server management, and workload management solutions by offering granular, real-time insights at the device level. Unlike traditional monitoring tools, AMI DCM creates a holistic view of the data center, ensuring actionable intelligence across power, cooling, firmware consistency, and device-level monitoring.

We demonstrated how AMI DCM can help a medium-sized data center (5,000 servers, 250 racks, PUE of 1.5) generate annual savings of nearly $2.2 million per year through just five use cases.

These five use cases represent only a subset of the many benefits AMI DCM provides and demonstrate why AMI DCM is an essential tool for modern data centers, helping them reduce costs, optimize infrastructure, and meet long-term operational goals.

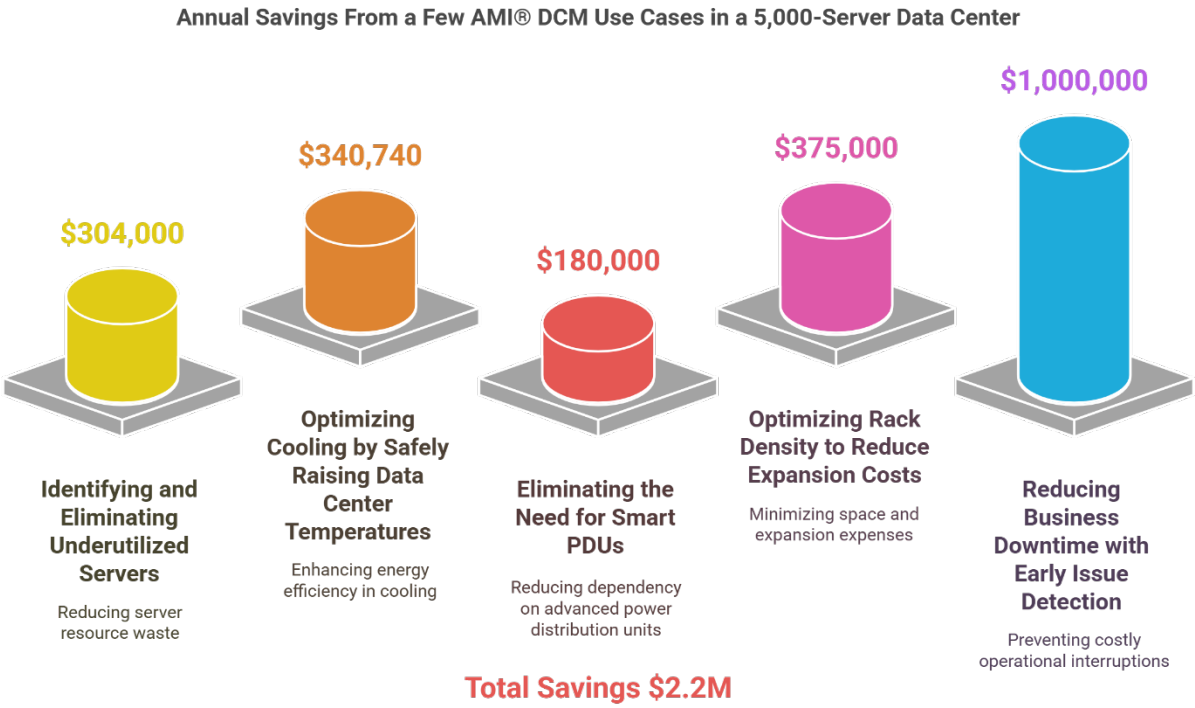


Figure 11: Annual savings from AMI DCM use cases in a medium sized data center.

References

- Deloitte. (2024). *Data Center Sustainability Insights*.
- Boyd Corporation. (2024). *Energy Consumption in Data Centers: Air versus Liquid Cooling*.
- EPRI. (2024). *Powering Intelligence: Analyzing Artificial Intelligence and Data Center Energy Consumption*.

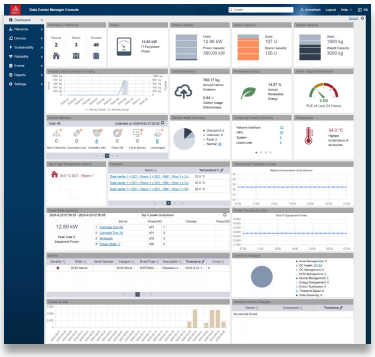
About AMI

AMI is Firmware Reimagined for modern computing. As a global leader in Dynamic Firmware for security, orchestration, and manageability solutions, AMI enables the world's compute platforms from on-premises to the cloud to the edge. AMI's industry-leading foundational technology and unwavering customer support have generated lasting partnerships and spurred innovation for some of the most prominent brands in the high-tech industry. For more information, visit www.ami.com.

ami® DCM
Data Center Manager

Scalable Data Center and AI Infrastructure Management

intel. AMD NVIDIA AMPERE.



What is AMI® DCM?

AMI® DCM is an advanced on-premise software for IT-centric data center management, supporting dense and heterogeneous environments.

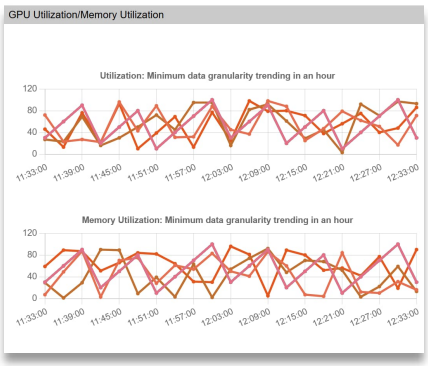
It manages servers based on different architectures and from different vendors, as well as networking, power, storage, and cooling devices.

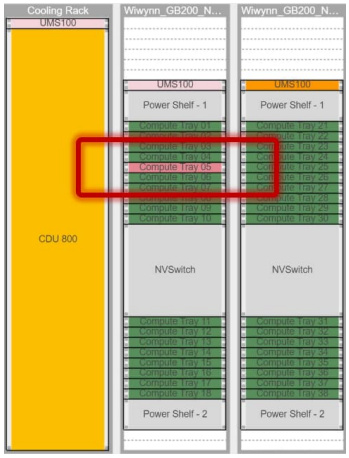
By collecting granular data on health, power, thermal performance, carbon emissions, utilization, and firmware, AMI® DCM delivers actionable insights that reduce capital end operating expenditures, enhance efficiency, and improve reliability – enabling data centers to meet their operational and sustainability goals.

GPU Management

AMI® DCM now supports NVIDIA data center GPUs, giving operators insights into GPU health, power, and performance and allowing them to perform GPU tasks and diagnostics.

This improves the reliability, uptime, resource utilization, and efficiency for AI and HPC clusters.





AI Infrastructure

AMI® DCM supports liquid-cooled NVIDIA® GB200 NVL72 racks with advanced monitoring and alerting.

It detects issues such as low coolant reservoir levels, leaks, and pump failures, delivering early warnings to prevent potential failures.

This prevents unplanned downtime, ensures peak AI workload performance, and prolongs hardware lifespan.

Latest Features

- ✓ NVIDIA GPU Management
- ✓ Arm CPU-Based Server Support
- ✓ Enhanced AMD-based server Support
- ✓ Realtime PUE Monitoring
- ✓ Liquid Cooling CDU Support
- ✓ Tiered Feature-Based Licensing

Key Benefits

- ✓ Enables granular and holistic IT device-centric data center management
- ✓ Delivers a single pane of glass web console for centralized management
- ✓ Provides silicon, platform, and vendor-agnostic server support
- ✓ Operates without agents, ensuring no impact on performance or security.
- ✓ Collects real-time power and thermal telemetry without additional hardware
- ✓ Monitors device health proactively to minimize unplanned downtime and extend device lifespan
- ✓ Optimizes rack density for improved energy efficiency and space utilization
- ✓ Produces granular thermal maps to identify inefficiencies and hotspots
- ✓ Streamlines firmware management

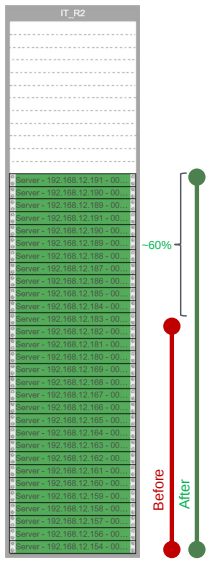
Device-Level Monitoring

Real-time, device-level energy and carbon emission monitoring and alerting ensures data centers can proactively reduce their carbon footprint and comply with regulations.

Power monitoring and planning can guide rack density optimizations, while server comparisons, consolidation, and replacement analysis improve energy efficiency and resource utilization, and guides procurement decisions.

Granular thermal mapping and cooling analysis detect hotspots and inefficiencies, enabling adjustments to lower cooling costs.

Power policies can also be applied to reduce power consumption for non-critical servers without impacting performance.



Cost Savings

Leveraging AMI® DCM's capabilities in a medium-sized data center can yield over $5M in CapEx and OpEx savings per year!

Energy & Sustainability

- ✓ Lowers energy consumption by identifying underutilized servers
- ✓ Supports server and rack-level power capping to optimize energy usage
- ✓ Monitors and manages carbon emissions to meet sustainability goals and regulatory compliance
- ✓ Reduces overcooling costs by advising temperature increases when possible

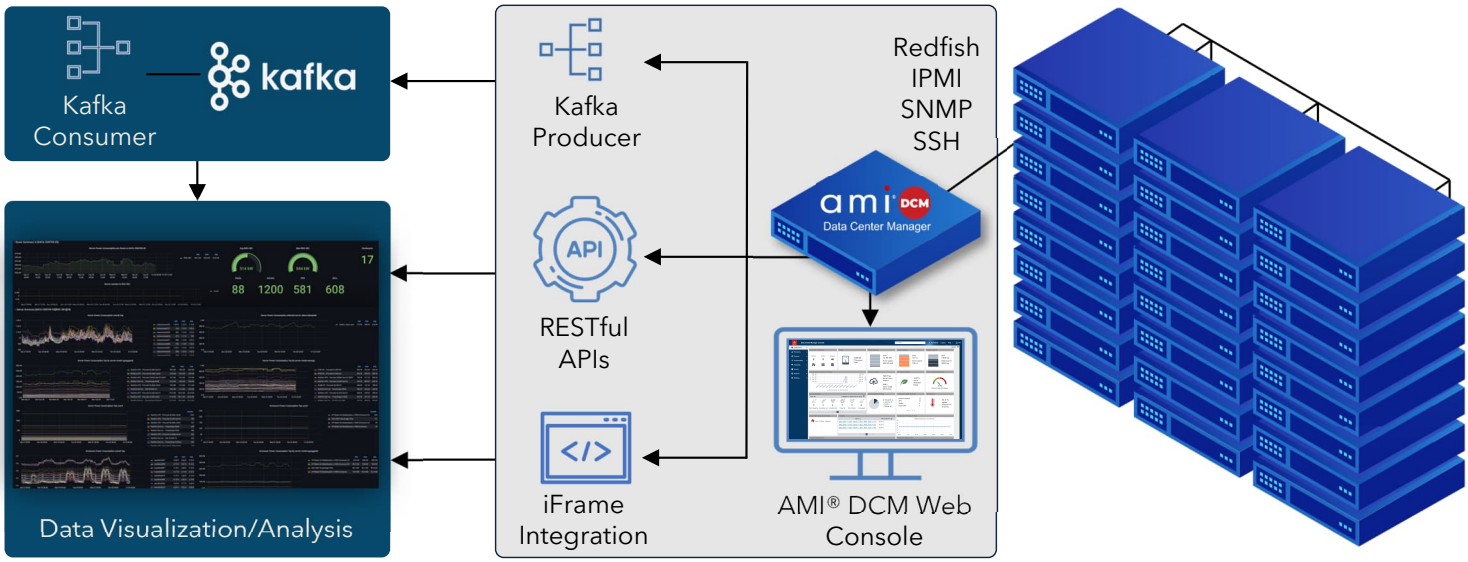
Supported Protocols

- ✓ Redfish
- ✓ IPMI
- ✓ SSH
- ✓ SNMP

For More Information

- ✓ Visit: www.ami.com/ami-dcm
- ✓ Contact: dcm_sales@ami.com

How AMI® DCM Works: Interfaces and Integration Points



Requirements	Details
Supported Operating Systems	• Microsoft Windows Server 2016, 1019, 2022, 2025 • Red Hat Enterprise Linux 8.10, 9.4 Server x86_64 • SUSE Linux Enterprise Server 12 SP5 and 15 SP6 x86_64 • Ubuntu Server 18.04.6, 20.04.6, 22.04.4, 24.04 x86_64 • Debian 11.10, 12.6 x86_64
Hardware Requirements	Recommended configuration for a scaled environment (e.g. managing up to 60,000 nodes) • 2 x 16 Core x86 Processor @ 2.60 GHz or higher • 192 GB RAM • 2 TB SSD • 10 Gigabit Network