

Using Intel® Intelligent Power Node Manager to Minimize the Infrastructure Impact of the EMC* Atmos* Cloud Storage Appliances

An EMC/Intel Proof of Concept

Abstract: Recent advances in networking technology are making it practical to deliver software and IT infrastructure as a set of policy-driven distributed services. This network of services is collectively known as the cloud. EMC Atmos* is a storage offering in this space, defined as Cloud Optimized Storage or COS, architected as massively scalable, cost efficient and service level-driven cloud storage solution. EMC Atmos runs on low-cost, high-density hardware delivered in an industry-standard form factor. One of the cost vectors under exploration is power and energy efficiency. In this report we report on experiments geared toward minimizing the burden of EMC Atmos appliances on the data center infrastructure by enabling the EMC Atmos application to set targets for the appliance's power consumption and make the appropriate tradeoff of power consumption against performance.*

Table of Contents

<i>Using Intel® Intelligent Power Node Manager to Minimize the Infrastructure Impact of the EMC* Atmos* Cloud Storage Appliances</i>	<i>1</i>
<i>Introduction</i>	<i>2</i>
<i>The Atmos Appliance</i>	<i>2</i>
<i>Need for Power Management in Cloud Storage</i>	<i>3</i>
<i>Available Power Management Technologies.....</i>	<i>3</i>
<i>Use Cases for Power Management</i>	<i>5</i>
<i>Cloud Plug-in Approach for Rapid Integration</i>	<i>6</i>
<i>A Power Management Proof of Concept</i>	<i>8</i>
<i>System Description.....</i>	<i>9</i>
<i>Power Management Experiments</i>	<i>11</i>
<i>Conclusions</i>	<i>14</i>
<i>Author</i>	<i>16</i>

*Enrique Castro-Leon, Ph.D.
End User Platform Integration
Intel Corporation*

*Ed Minasian
Principal Product Manager
EMC Cloud Infrastructure Group*

Rev 1.3

January 2010

Introduction

Recent advances in networking technology are making it practical to deliver software and IT infrastructure as a set of distributed services. This network of services is collectively known as the *cloud*.

The technological impact of the cloud to customers and IT providers is significant, but perhaps of more consequence will be on the business side in the way IT is delivered and in changes of patterns of the ownership of data, applications and infrastructure and the services wrapped around them.

For instance, one way end-users will utilize cloud computing is to access their applications and information from a third-party provider such as a large telecommunications company with the resources and clout to build a global cloud infrastructure. That cloud infrastructure will make massive amounts of unstructured information available on the Web, and will require a policy-based approach to efficiently disperse the information worldwide.

Existing storage technologies, however, were not designed to deliver cloud services, and so cloud service providers require a new category of storage that can address their requirements, namely:

Massive scalability. Providers must assemble applications and information from multiple sources to create unique experiences for the consumer.

Global distribution. An information policy is necessary to enable the applications and information to move closer to the consumer.

Efficiency at scale. Providers need an easy, cost-effective way to operate and manage massive amounts of unstructured information.

This new category of storage is defined as Cloud Optimized Storage (COS). COS complements and optimizes the cost of traditional storage categories. COS brings resources, tools and processes for delivering IT services to customers more efficiently.

As part of EMC's COS offering, EMC* Atmos* implements a multi-petabyte information storage and distribution capability. It combines massive scalability with service level-driven automated data placement to efficiently deliver content and information services anywhere in the world. EMC Atmos allows managing the geographic dispersion of data to attain any desired level of reliability. EMC Atmos supports the policy-based approach mentioned above, providing a solid foundation from which to build cloud storage.

EMC Atmos operates logically as a single entity, yet its use of metadata and sophisticated policy engine facilitates easy mapping to business policy and helps support service level agreement (SLA) transparently to users. These qualities combine to increase operational efficiency, reduce management complexity and cost of operations.

Cloud technologies represent an emerging IT architecture: one with greater scalability, elasticity, and lower costs. An effective cloud strategy begins with establishing a solid cloud infrastructure from which to build, virtualize and deploy services and applications. The storage component is no exception to an effective cloud strategy.

The Atmos Appliance

EMC Atmos software runs on low-cost, high-density hardware that is delivered in an industry standard form factor. The hardware is customer-installable, serviceable, and configurable based on capacity and/or compute requirements.



Figure 1. EMC Atmos Appliance.

EMC Atmos is housed in a modular appliance consisting of an EIA standard 40U or 44U cabinet deployed with six to sixteen Intel® Xeon® 5500 Series processor-based servers. Each server is connected to direct-attached storage devices (DASD) housed in disk array enclosures (DAEs). Each DAE can house up to fifteen SATA drives. Each server is connected to one to four DAEs depending on system configuration with SAS wide port (4 lanes) cable connections.

Connectivity from an EMC Atmos appliance to the customer's network is achieved through Gigabit or 10 Gigabit Ethernet.

Need for Power Management in Cloud Storage

EMC Atmos is currently supported with three appliance configurations: the WS1-120, WS1-240 and WS1-360 with 120, 240 and 360 hard drives respectively. The appliances have a nominal maximum power consumption of 4.9, 10.0 and 10.3 KW. Each appliance is designed with redundant power through standard NEMA L6-30, 220 V plugs providing at least 1+1 redundancy at the cabinet level as well as at the chassis (server, DAE, switch) sub-power supply level.

Power efficiency represents a valuable feature to EMC Atmos customers from the standpoint of minimizing operating expenses. Even though cloud storage represents an evolution from traditional data center-based, custom-designed storage, the need for power efficiency remains.

In fact, the ROI from making EMC Atmos appliances power efficient is very high because benefits from this effort are multiplied over all the units deployed, whereas efforts to optimize an in-house data center are by their very nature one-of-a-kind. This added efficiency is attained on top of the inherent efficiency from cloud resource pooling: multi-tenancy implies that the power consumed is shared over the users of a cloud device with lower power consumption compared to dedicated devices.

From a power consumption perspective, the servers use Intel® Xeon® 5500 Series processors representing the state of the art in power efficiency. Some of the power efficient features require no particular action from the user to be elicited, for instance low idle power consumption under Intel® SpeedStep® technology and Demand-Based Switching.

Available Power Management Technologies

In addition to passive mechanisms such as the Intel® SpeedStep® technology, which require no specific user action other than perhaps changing a BIOS setting or a system configuration parameter, a number of active power management technologies have been incorporated into Intel-based servers. Active power control mechanisms have a broader operating envelope, over passive mechanisms, and being under direct

EMC* Atmos Power Management Proof of Concept

application control, their effect can be targeted much more precisely to maximize benefits and minimize side reactions.

These technologies are accessible through industry-standard interfaces such as the Intelligent Power Management Interface* (IPMI) and Web services and hence available to any hardware or software integrator or service integrator to incorporate into an application.

We can look at the server equipment in a data center as a set of nested abstractions, not unlike the well known Russian matryoshka dolls are put together: Chipsets are used to bind together the CPUs and the memory in a server. A server carries direct-attached storage devices (DASD). Servers are organized in racks, and racks are organized in rows. The aggregation of rows encompasses all the servers in a data center. These relationships are illustrated in Figure 2. Opportunities exist to reduce power consumption at every level.

Today our main lever for power control is throttling the power consumed by the CPUs up and down. In the near future we can expect to see memory power control added to the mix. The basic mechanism of CPU voltage and frequency scaling allows moving the power consumption of a CPU by a few tens of watts up or down. Aggregating this capability over the tens of thousands of CPUs in a data center expands the range of attainable power control to tens of kilowatts or even hundreds.

There is also a potential synergistic effect captured by the PUE (power usage effectiveness) factor as defined by The Green Grid¹ industry group. If servers use less power, the power allocated to the cooling equipment can be ratcheted down as well.

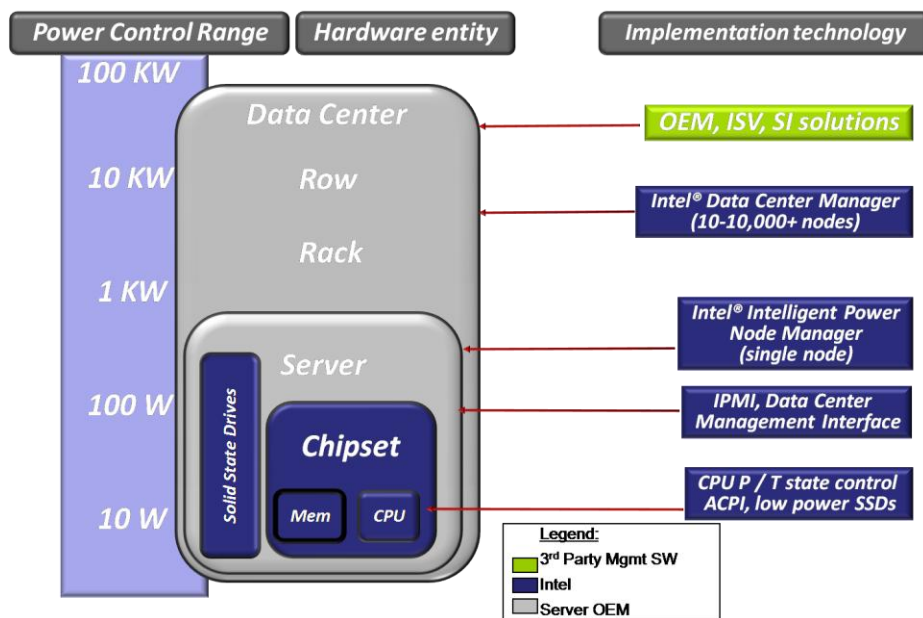


Figure 2. Power control hierarchy.

Power control for groups of servers is attained by composing power control capabilities of power control of each server. Likewise, power control for a server is attained by composing CPU power control as illustrated in Figure 2.

The power control capability through voltage and frequency scaling is calibrated into actual watts consumed by a server through the Intel® Intelligent Power Node Manager firmware that runs in server chipsets using instrumented power supplies for actual power readouts. The instrumentation follows the PMbus² standard. Readouts for target power numbers and actual power consumed are performed through IPMI-formatted messages in the IP-based management network. This network can be the same as the application network, or isolated into a management network for added security.

¹ The Green Grid consortium, <http://www.thegreengrid.org/>.

² PMBus, <http://pmbus.org/> is an initiative of SMIF, Inc., the System Management Interface Forum.

EMC* Atmos Power Management Proof of Concept

Conceptually, power control for thousands of servers in a data center is implemented by aggregation through a series of coordinated set of nested mechanisms. The actual mechanisms are application-specific. For instance, the Intel® Data Center Manager software can be used to aggregate multiple servers into logical groups.

A more detailed description of the power management architecture for server equipment can be found in the Intel Server Room article³ *A Reference Architecture for Cloud Storage Power Management*.

Use Cases for Power Management

The power control capabilities provided by Intel Intelligent Power Node Manager at the single server (node) level and Intel® Data Center Manager at the server group level, open up a number of potential uses as shown in Table 1.

Table 1. Use cases for Intel Intelligent Power Node Manager and Intel Data Center Manager

Use Case	Name	Capability
1	Node power history	Generate historical power consumption records per node
2	Compute subsystem power history	Generate historical power consumption aggregated over nodes in an appliance
3	Appliance power history	Generate power consumption history for EMC Atmos appliance
4	Node power control	Define server power consumption targets
5	Compute subsystem power control	Set up server power consumption targets for servers in an appliance as a group
6	Single node event monitoring	Enable power policy engine in appliance to monitor and take action on specific power events: capping over power thresholds or over temperature events
7	Group event monitoring	Apply event handling to logical groups of nodes
8	Compute subsystem global cap, local policy	Establish management policies by sub-groups for servers in an EMC Atmos appliance
9	Appliance global cap, local policy	Establish management policies for EMC Atmos appliance
10	Increase rack density	Use historical power consumption or power capping to maximize number of nodes in a rack

³ <http://communities.intel.com/docs/DOC-4316>

11	Power shedding	Turn off or reallocate/minimize power consumption in response to emergency
12	Power forecasting	Use Intel Data Center Manager historical data for planning purposes
13	Fine grained power monitoring/chargeback	Per node or per tenant power billing

Intel Data Center Manager builds a data base of all the sampled power readings from the instrumented power supplies. This information can be queried through the API available from the SDK. Hence use cases 1, 2 and 3 can be implemented without undue effort using built-in Intel Data Center Manager capabilities.

Use case 4 is a basic power management capability implemented by Intel Intelligent Power Node manager and can be implemented through the appropriate IPMI message or through a call to Intel Data Center Manager. The aggregation described in use case 5 is implemented by Intel Data Center Manager.

Use cases 6 and 7 represent the event handling facility supported by Intel Data Center Manager.

Use cases 8, 9 and 10 reflect the capability of Intel Data Center Manager to enforce power quotas to sub-groups within a larger group.

Use case 11 implements what essentially amounts to a controlled crash during impaired operations. It is applicable during emergencies where power usage needs to be minimized at all costs. Specific servers, presumably running lower priority or recoverable workloads can be turned off through specific IPMI messages based on a pre-arranged policy. The remaining servers are run at an aggressive power cap until it is safe to shut them down.

Applying statistical methods to power usage records over time enables power forecasting as per use case 12 for planning purposes or even to optimize operations to the extent that workloads follow predictable patterns. If power usage per server can be correlated to specific tenants in an appliance, it would be possible to break down power consumption on a per tenant basis to implement fine-grained power metering.

Cloud Plug-in Approach for Rapid Integration

The power management technologies available in Intel-based servers would be of little use if they were difficult to integrate. These technologies were architected with a modular approach in mind allowing system integrators to pick the optimal tradeoff between level of abstraction and the granularity of control.

The implementation of power policies at the IPMI level allows very fine grained control down to the individual machine. This approach is useful under particular circumstances, but not as a general method for application power control due to the labor involved. The situation is similar to using assembly language in software development projects.

Intel Data Center Manager as a software development kit (SDK) can be used as a ready-made module with a number of power management capabilities already built in such as the support of logical server groups, eventing and historical records. Hence this SDK can be incorporated as a power management module as a retrofit to an existing application. The API is supported through a Web services interface, with the technology's inherent advantages of late binding and rapid integration.

The application implements the desired power management strategy and makes the appropriate calls to the Intel Data Center Manager to carry out that strategy. Intel Data Center Manager also provides a small-scale reference user interface that allows implementing, testing and debugging power management capabilities with incremental development enabling progressive refinement of the control policies while minimizing the risk of large scale rework from deferring testing toward the end of the project.

EMC and Intel undertook a proof of concept project to endow the EMC Atmos application with a power management capability in a resource-constrained exercise. This project is still in progress at the time of writing.

In Figure 3 we have integrated all the technology components discussed so far into an abstract cloud storage power management architecture. This architecture supports the use cases mentioned above as well. The numbered paragraphs below correspond the numbered tags in Figure 3.

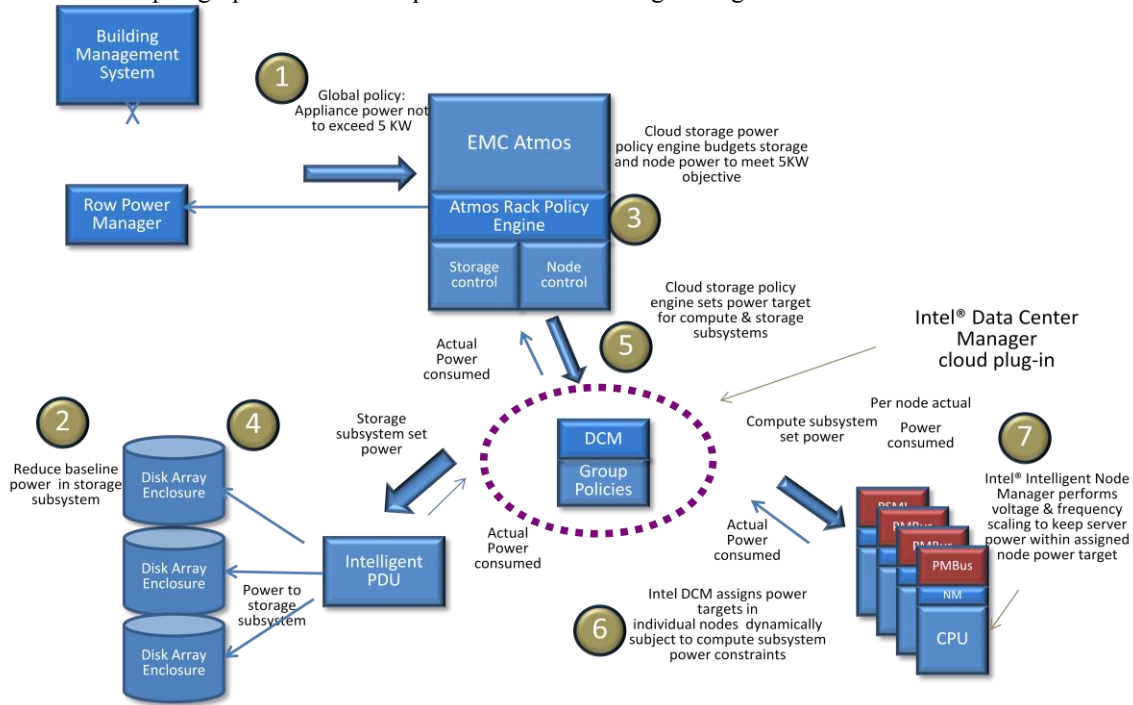


Figure 3. Power Management Cloud Plug-in Architecture Applied to EMC Atmos.

1. The cloud storage application may be implemented with multiple physical Atmos appliances arranged in a row. We postulate a management application regulating the power behaviors of the appliances in that particular row, the *row power manager*. The row power manager may be connected to the *building management system* (BMS) application, whose role, among others, is to oversee power management across the whole data center.

The row power manager implements a number of policies that get mapped into specific directives to each storage appliance. For instance, the row power manager may instruct the EMC Atmos appliances under its command to operate in a power constrained mode imposing a guard rail in power consumption not to be exceeded by the appliance. Without this guard rail, the circuit feed for the row of appliance would have to be provisioned for the worst case assuming concurrent peak consumption in all the appliances in the row. This condition is rarely reached, if ever. Because this power needs to be allocated, but rarely, if ever reached, it effectively represents stranded capacity. The stranded capacity increases data center capital expenses, avoidable through improved management practices. This application of power control technology is an instance of use cases 4 and 5, node control and compute subsystem power control described in Table 1. We call it *branch circuit optimization*.

2. In addition to the application of power management-specific technologies, there are emerging technologies that bring reduced power consumption. A case in point is the replacement of hard drives with solid state drives (SSDs). SSDs consume less than 1 watt at idle and less than 5 watts vs. 10 to 15 watts typical of mechanical hard drives. Hence using SSDs will yield a lower $P_{baseline}$.

EMC* Atmos Power Management Proof of Concept

3. The rack policy engine in the storage appliance oversees its power consumption by monitoring the power draw from the power distribution unit (PDU) feeding the hard drives and the power consumption by the server subsystem as reported by the instance of Intel® Data Center Manager regulating the power consumption of the servers.
4. The implementation of the storage appliance may provide a monitor-only capability for the storage subsystem, in which case the appliance policy engine needs to meet the power quota for the appliance by regulating the power consumed by the servers in the appliance.
5. The rack policy engine in the storage application assigns a power target to Intel® Data Center Manager. This power target can change dynamically depending on workload conditions and the policies set at the higher levels.
6. Intel® Data Center Manager takes the overall power quota for the server subsystem and divides it across the servers in the appliance.
7. Intel® Intelligent Power Node Manager instances adjust CPU frequency and voltage scaling accordingly to meet the quota imposed by Intel® Data Center Manager.

In our experience, interfacing an application to the Intel® Data Center Manager usually takes less than a weeks' time even with implementation team not previously exposed to the API; all it takes is a few Web services calls. Most of the effort goes into validating the new capabilities. Because of the small effort involved in the interfacing, Intel® Data Center Manager, for practical purposes, functions as a *plug-in* module to quickly add a power management capability to a cloud-based application; in this case, a cloud storage application. This capability is added without need of re-architecting the original application in any fundamental way.

A Power Management Proof of Concept

A joint EMC/Intel team put together a demonstration system shown at the 2009 Intel Developer Forum in San Francisco. This occasion presented the first opportunity to assemble working end-to-end system within the context of the Atmos PoC currently under execution. We report on experiments related to an implementation of branch circuit optimization using the plug-in architecture described above.

One of the goals for operating a set of appliances under a power constrained regime is to prevent breaker outages. Breaker outage prevention would be a last resort measure. A more common situation will be the use of power constraints to meet SLAs, consistent with the design goals of EMC Atmos. Meeting the SLA may require the use of N+1 redundancy, for instance feeding a group of appliances with three branch circuits of the same rating, with only two needed for normal operation.

Hence, if we impose a power cap to the power available from two of the three branch circuits, the group of appliances can continue with normal operation even if one of the branch circuits fails. If a failure occurs, redundancy is lost. In this case the imposed power cap prevents the appliances from drawing more power than the amount possible with two branch circuits and triggering a breaker outage.

In Figure 4 we start with an unconstrained system at idle up to T_0 . Workload ramps up until the power draw exceeds P_{max} . Most of the time the system will operate under the power ceiling such as the interval between T_2 and T_3 , with only occasional excursions (T_1 to T_2 , T_3 to T_4)

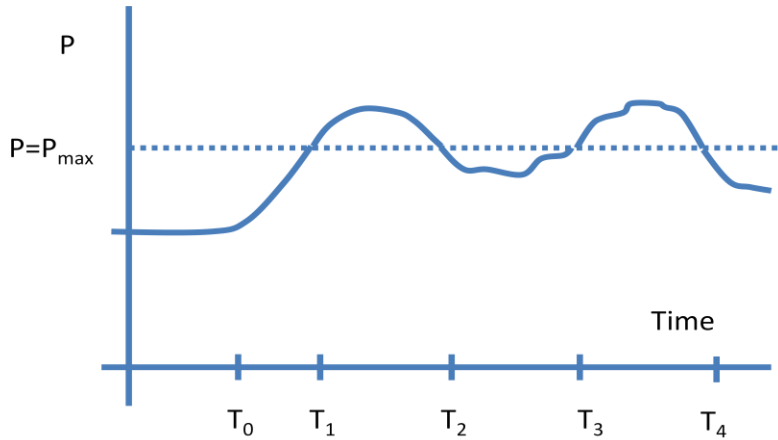


Figure 4. Unconstrained power consumption.

Figure 5 depicts operation under a power capped regime where Intel Intelligent Power Node Manager is used as a guard rail mechanism to keep power consumption at or below the P_{max} boundary to maintain $N + 1$ power supply redundancy, allowing the system to operate without exceeding the P_{max} envelope at all times.

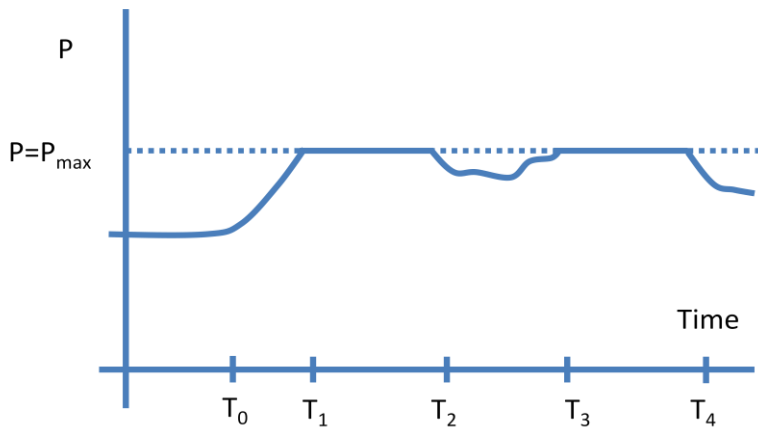


Figure 5. Operation under a power constrained regime.

Unlike what Figure 5 suggests, the system will likely operate under the P_{max} cap most of the time, hitting the ceiling infrequently, if ever. Operation under a power constrained regime may have an effect on CPU performance. Since capping events are infrequent, users will not experience performance degradation, if at all.

Even if a second branch circuit fails, this event does not necessarily translate into catastrophic failure. It may be possible to bring the power cap to a maximum level, perhaps with selective shutdown of non-essential node to the point that the system can continue operations with a single branch circuit, albeit with degraded performance.

Actual power plots from the Intel Data Center Manager reference GUI are shown in the next figures based on the conceptual view just described.

SYSTEM DESCRIPTION

Because of the limited time and engineering resources available for the demo, it was implemented to run on top of VMware* VSphere* 4 and hosted on the smallest hardware footprint as shown in Figure 6. For DASD storage, instead of the usual SAS disk array enclosures we used only hard drives in the drive bays in two Intel® Xeon® 5500 Series servers. Also it was not practical to integrate an adjustable I/O workload in

EMC* Atmos Power Management Proof of Concept

the short time available, and hence a synthetic workload was used to simulate the ups and downs of server workload.

One of the two servers hosted the EMC Atmos software, running in four virtual machines. A fifth virtual machine was loaded with Microsoft* Windows Server 2008 running the Intel® MaxPwr synthetic workload used to impose a background workload on the system.

The second server ran Microsoft* Windows Media Server. EMC Atmos was presented to Windows as a single CIFS file containing a number of videos to be streamed.

The implementation efforts were focused in delivering a functionally correct demo without an attempt to optimize performance. Even then, the system was able to support the delivery of seven HD video streams without hiccups. Power capping did not have any appreciable effect on the video frame rate.

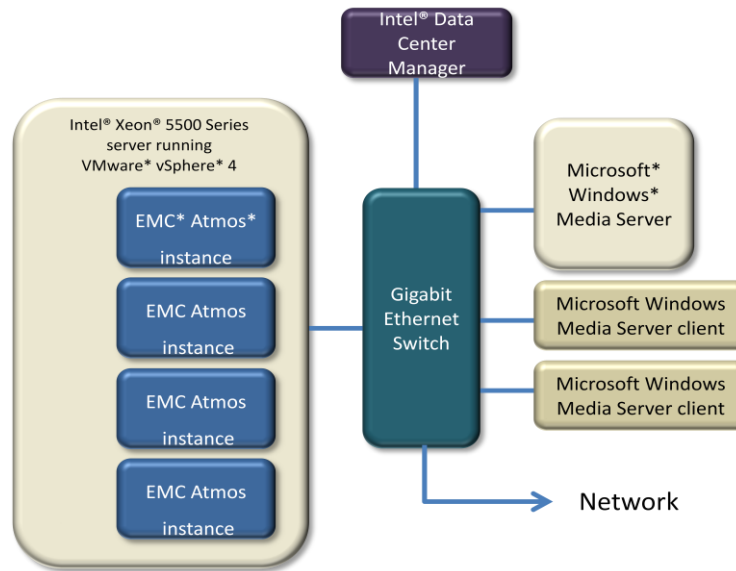


Figure 6. Hardware Setup.

The hardware diagram actually obscures the logical simplicity of the test rig, whose logical diagram is shown in Figure 7. Even though there are multiple instances of EMC Atmos, it operates as a single logical entity as mentioned in the beginning, functioning as a cloud-based data source for Microsoft* Windows* Media Server. Also note that the Intel Data Center Manager SDK becomes an entity embedded within Atmos, acting as the single proxy for Intel Intelligent Power Node Manager to carry the EMC Atmos power policies.

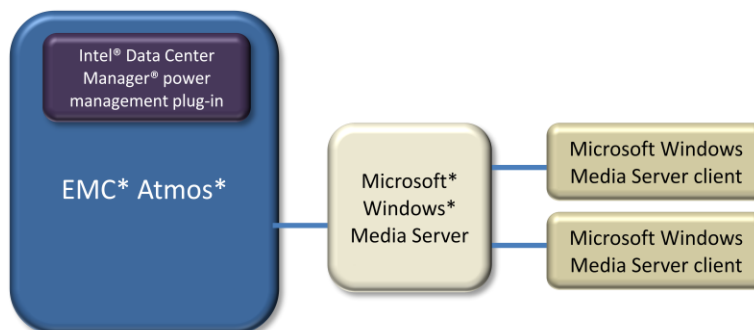


Figure 7. EMC Atmos Microsoft Windows Media Server Application Logical Setup.

POWER MANAGEMENT EXPERIMENTS

Figure 8 shows the actual power consumption trace starting with the host powered up and idle. The power consumption numbers correspond to the single physical host running EMC Atmos using the hard drive bays in the server itself. Hence power consumption numbers are considerably lower than those in a full size appliance.

We ran the server for few minutes until it stabilized before booting the virtual machines and starting Atmos, and give it a few more minutes until it stabilizes again. Note that the booting VMs and starting Atmos induces a significant power bump. This behavior suggest another potential use case in data centers with large numbers of servers whereby servers are started in staggered groups to manage power draw during startup.



Figure 8. Virtual Machine System Initialization.

As shown in Figure 7, idle power is about 210 watts. With the VMs and EMC Atmos power consumption increases to about 220 watts. Power consumption is somewhat unsettled, probably due to interactions between VMs, EMC Atmos and the application server.

EMC* Atmos Power Management Proof of Concept

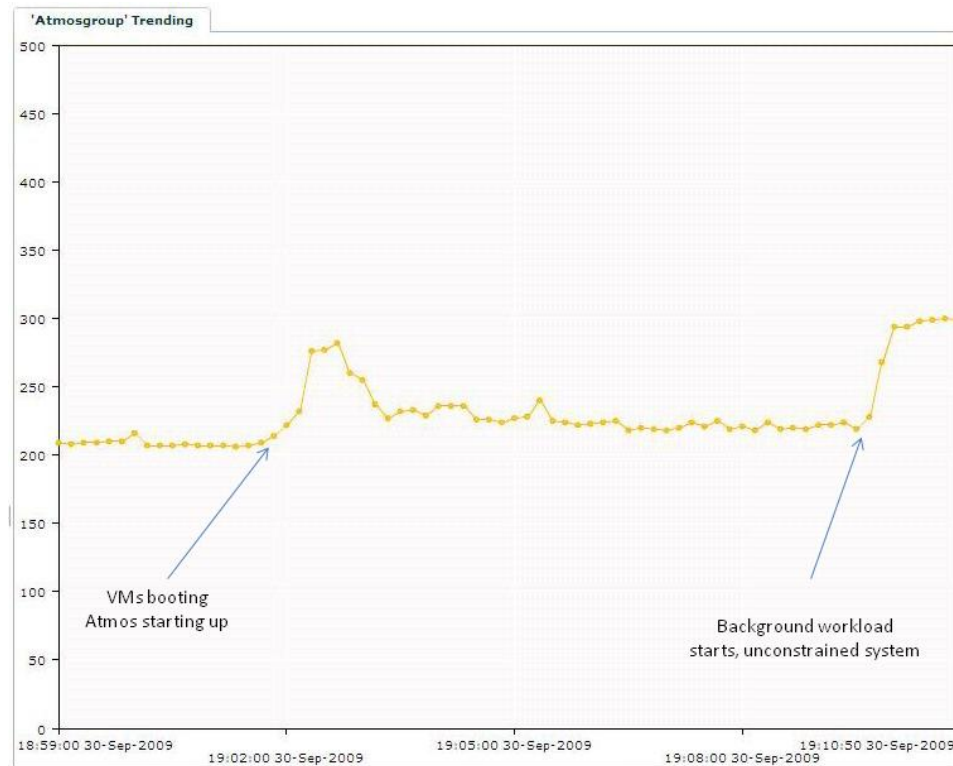


Figure 9. Unconstrained workload.

Figure 9 shows one run throttling the background workload to 100% to test the power capping range. The run is started after the system reaches steady state with EMC Atmos. The graph shows an appreciable power proportional computing range from about 220 watts to 310 watts as shown in Figure 8.

As shown in Figure 10, we remove the workload a couple of minutes later and as expected, the system settles back to the previous baseline.

In Figure 10, toward the right we impose an aggressive cap of 250 watts and throttle once more to 100%. The system complies, somewhat reluctantly as the presence overshoot indicates.

EMC* Atmos Power Management Proof of Concept

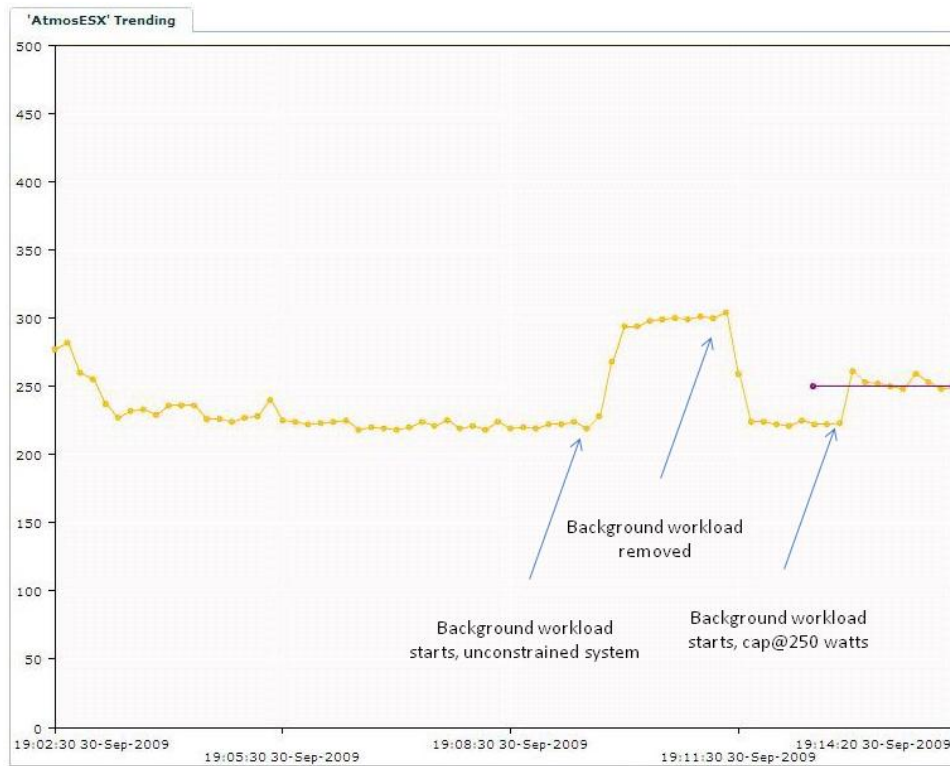


Figure 10. Unconstrained operation and start of power capped regime.

In Figure 11 toward the right edge we repeat the same process with a more generous cap of 270 watts. Note that this time Intel Data Center Manager maintains a smoother capping action with less overshoot.

Once more, we remove the workload and the system returns to baseline EMC Atmos power consumption.

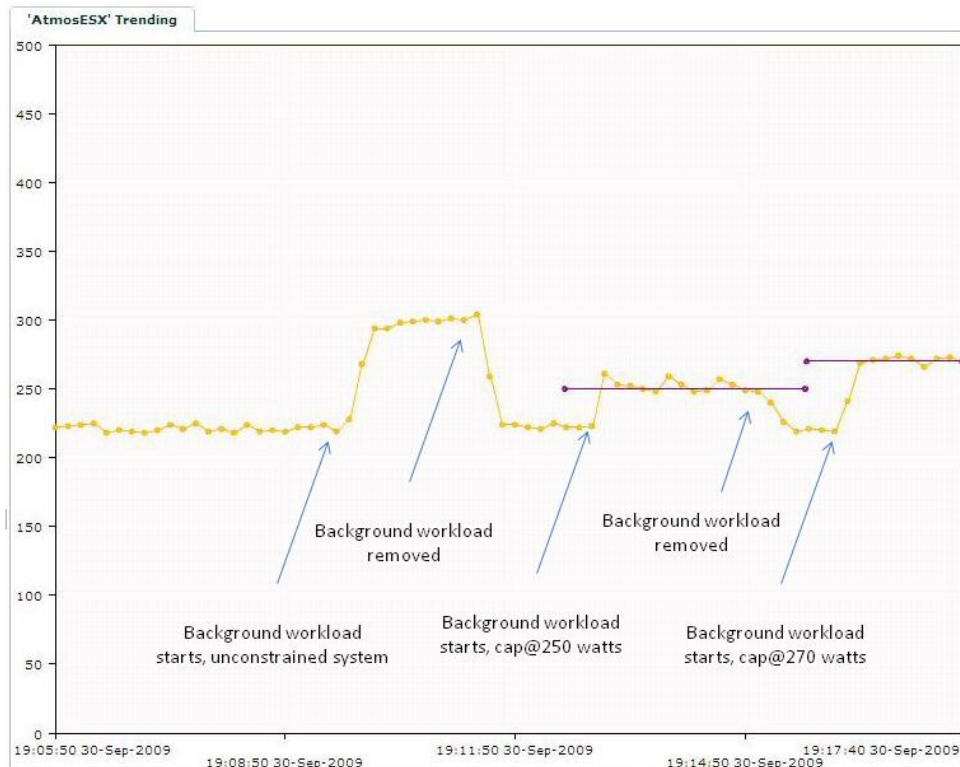
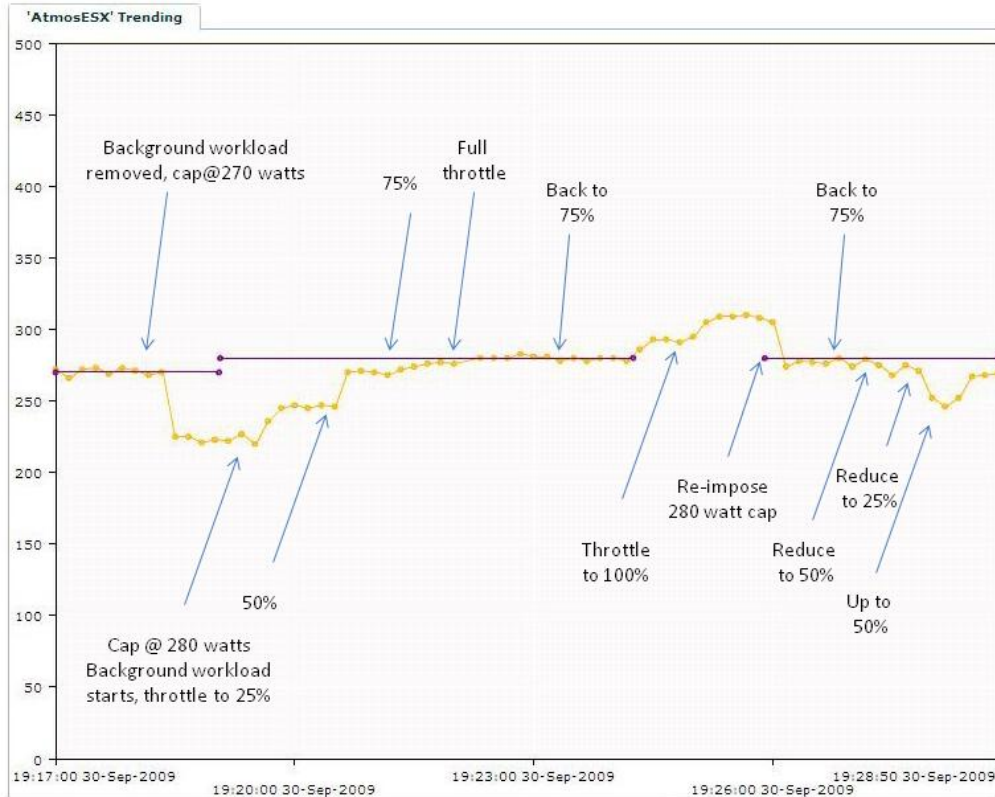


Figure 11. Power caps set at 250 and 270 watts.

In the next experiment, shown in Figure 12 we simulate a variable workload using 25% increments. This time we use a 280 watt cap. Note that this ceiling is reached at the 75% level. Pushing for full throttle barely buds the power consumption as Intel Data Center Manager clamps it down.

At this point we throttle back to 75% and then remove the cap. Note how power consumption, now unconstrained, spills over to 280 watts. Increasing the workload to 100% increases consumption to 310 watts, the limit we saw before and well beyond the 280 watt boundary. This behavior illustrates the role Intel Intelligent Power Node Manager and Intel Data Center Manager plays in keeping power consumption clamped at a pre-determined boundary and the spillover that takes place if for some reason the cap is removed.

**Figure 12. Power “spills” over the gap when the power cap is removed.**

Conclusions

The experiments described in this test report were performed with a human observing the system behavior through the Intel Data Center Manager reference GUI and setting power policies manually. A next logical step will be to integrate power policies with the EMC Atmos application with EMC Atmos driving the policies through calls to the Intel Data Center Manager API. The test is to be carried out in a production-size EMC Atmos appliance with a full complement of nodes and disk array enclosures.

The experiments with the full appliance would include A/B testing, that is, demonstrating that with power capping on the power consumption of an Atmos appliance stays within the preset limit, and to demonstrate the existence of significant power excursions when power capping is off. A number of methods can be used to inject a disturbance or provide a stimulus. One method would be through a synthetic workload as in this experiment. Results will be more realistic if we apply disturbances associated with the operation of EMC Atmos, for instance by imposing workloads associated with the encryption or decryption of data

EMC* Atmos Power Management Proof of Concept

streams, XML processing and data compression and decompression, preferably using the Grinder workload generator.

This test setup environment would allow testing the feasibility of regulating the power consumption of an appliance by regulating the power consumption of the nodes within. Beyond the verification of the branch circuit optimization use case the additional experiments will provide a foundations for more complex use cases that implement extreme power/workload scalability. These test cases would include controlled server shutdowns and hard disk drive spin-downs.

A desirable target to attain is to bring idle power consumption from the 50% attainable with a single Intel Xeon 5500 Series server down to 15% for the whole appliance. Ideally power scalability should enable bringing down power consumption close to the expected load average.

Author

Enrique Castro-Leon is an enterprise architect with End User Platform Integration at Intel Corporation.

This paper is for informational purposes only. THIS DOCUMENT IS PROVIDED "AS IS" WITH NO WARRANTIES WHATSOEVER, INCLUDING ANY WARRANTY OF MERCHANTABILITY, NONINFRINGEMENT, FITNESS FOR ANY PARTICULAR PURPOSE, OR ANY WARRANTY OTHERWISE ARISING OUT OF ANY PROPOSAL, SPECIFICATION OR SAMPLE. Intel disclaims all liability, including liability for infringement of any proprietary rights, relating to use of information in this specification. No license, express or implied, by estoppel or otherwise, to any intellectual property rights is granted herein.

Intel, the Intel logo, and the Intel Leap ahead logo, are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

* Other names and brands may be claimed as the property of others.

Copyright © 2009, Intel Corporation. All rights reserved.

