# The IBM System x iDataPlex dx360 M4: Superior Energy Efficiency and Total Cost of Ownership for Petascale Technical Computing

*A Total Cost of Ownership (TCO) Study comparing the IBM iDataPlex dx360 M4 (with Intel Xeon E5-2600 processors – Sandy Bridge) with traditional x86 (Intel Xeon 5600 processors – Westmere EP) clustered rack systems for High Performance Technical Computing (HPC)*

**Srini Chari, Ph.D., MBA** **Sponsored by IBM**

**March, 2012**

chari@cabotpartners.com



## Executive Summary

*Energy and power consumption are the topmost challenges in the race to Exascale computing[1]. Other constraints include memory, storage, networks, resiliency, software and scalability. Power limits the total number of components that can be packaged on a chip and the total energy required by Petascale/Exascale performance capable systems restricts their location to places closer to sources of affordable power. Extrapolating from current power consumption rates from the Top500 and Green500 lists, Exascale power requirements are of the order of Gigawatts, large enough to power multiple modern cities. Further, escalation of global electricity costs, paucity in data center floor space, and complexity of manageability pose additional challenges for HPC.*

*High performance supercomputing clusters and parallel systems constitute the largest installed base for supercomputers today[2]. The most prevalent architectures for high performance supercomputing in the Petascale to Exascale range are: traditional microprocessor (e.g. x86 CPUs) based multi-core, multi-socket rack based clusters, hybrid cluster racks with a mix of x86 CPUs and Graphics Processing Units (GPUs), the latest IBM System x – iDataPlex dx360 M4 servers based on the Intel Xeon E5-2600 processors (latest generation of high performance x86 processors), and other exotic supercomputing architectures.*

*The key strengths of pure x86 clusters are standardized components and slightly lower acquisition costs. However, when scaling to a few Petaflops, these systems have significantly higher energy and Site Infrastructure costs compared to iDataPlex dx360 M4 clusters making the TCO for the iDataPlex very attractive – our analysis indicates that this could be as much as 57% lower for the iDataPlex dx360 M4 system as compared to typical x86-based cluster in the Petaflops range. And this TCO advantage gap becomes even more pronounced for larger clusters. Here, we do not consider GPU based hybrid systems as they bring in additional challenges of wider acceptance by the HPC community and software migration complexity and costs, reliability and availability issues that add to the overall TCO. In terms of energy efficiency, scalability, and overall TCO, IBM clearly leads the pack. Our analysis shows that iDataPlex dx360 M4 servers have an edge over the typical x86-based HPC server clusters.*

*In this paper, we detail our TCO methodology and analysis and discuss our results obtained when we compare standard Westmere based cluster systems versus the new iDataPlex dx360 M4 system. Data for anchor systems selected for the study were sourced from public data available for existing supercomputing systems. A TCO model was created for each type of architecture using the Uptime Institute's data center TCO calculator as the base and then customized to HPC environments. This was enhanced with findings of our earlier analysis[3] to account for RAS costs in supercomputing clusters. Data from anchor systems were fed into the enhanced calculator and results analyzed to arrive at comparative insights that clearly indicate iDataPlex dx360 M4 as the most promising, energy and cost effective solution for the Petascale supercomputing needs in the x86 cluster systems market.*

Cabot Partners
Optimizing Business Value

[1] Darpa Study identifies four challenges for Exascale computing: http://www.er.doe.gov/ascr/Research/CS/DARPA%20exascale%20-%20hardware%20(2008).pdf
[2] Horst D. Simon: Petascale systems in the US http://acts.nersc.gov/events/Workshop2006/slides/Simon.pdf
[3] Why IBM systems lead in Performance, Reliability, Availability and Serviceability (RAS): http://www-03.ibm.com/systems/resources/systems_deepcomputing_IBMPower-HPC-RAS_Final-1.pdf

## Introduction

The high end technical high performance computing market continues to expand as the race to Exascale computing heats up across the globe. Today, there are many installed systems with performance of the order of 100s of Teraflops and a few Petaflops in the US, Europe, Japan and several other emerging countries such as China. The Top500 list, Green500 list and NERSC estimates indicate that sustained Petaflops systems have approximately 1.5 to 5 million cores[4]. These systems today consume power of the order of 20MW and require almost 16000 square feet of floor space with over $12 million per year as electricity costs.

Given the complexity and scale of supercomputing systems, it is important to analyze how different architectures are dealing with power and energy challenges. This paper is a realistic TCO analysis of the fastest x86-based supercomputers today, in a bid to support well-informed business and financial decisions when evaluating and deciding on various supercomputing systems available in the marketplace.

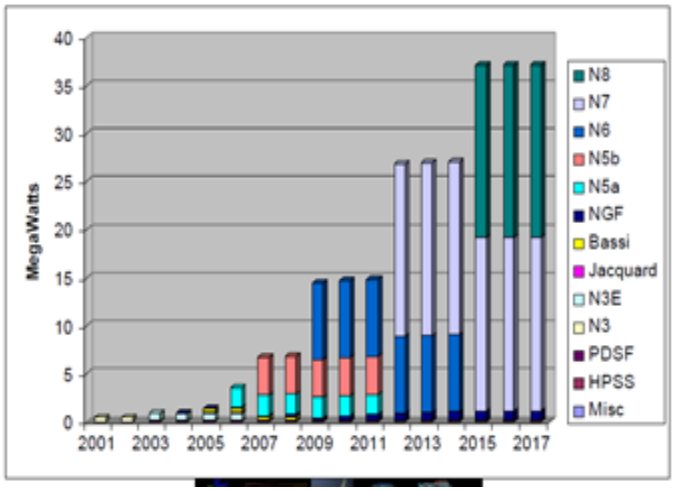


**Figure 1: Projections for Computing Systems Power without cooling (*Source: NERSC*)**

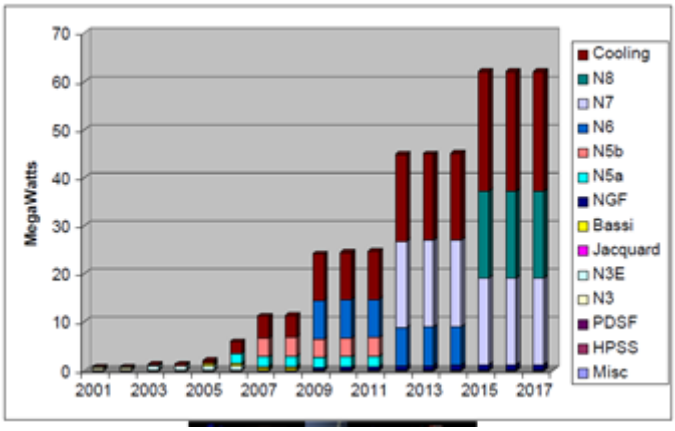


**Figure 2: Projections for Computer Room Power System + Cooling (Source: NERSC)**

The key supercomputing players today include IBM, HP, Dell, Cray, and SGI with other smaller vendors trying to get a foothold[5]. The available offerings span a broad range of architectures from pure x86 clusters to hybrid CPU-GPU ones, iDataPlex dx360 M4 servers and also to ultra-scalable systems such as IBM's Blue Gene/Q. IBM leads the market in very large scale supercomputing offerings for superior scalability, portfolio breath, increased energy-efficiency, compute density, operational efficiencies[5].

[4] Horst D. Simon: Petascale systems in the US http://acts.nersc.gov/events/Workshop2006/slides/Simon.pdf
[5] IDC Vendor HPC Market share: http://www.hpcuserforum.com/presentations/Tucson/new%20IDC%20market%20slides%204%20Supercomputer.ppt

One of the major attributes of a supercomputer is its processing performance. It is usually measured in total or peak Teraflops (TF). The performance of the highest ranking supercomputer in the Top 500 supercomputer list went from around 140TF in 2005 to an impressive and previously unimaginable 1 Petaflops (PF) in 2008. However, in recent years escalating energy costs have brought more focus on another metric in addition to pure flops. That is the metric of gigaflops/watt.

Today, the horizon of multi-petaflop computing poses the question: What are the true trends in the total costs that are incurred when making investments in supercomputers and what is the business value moving forward? In addition to acquisition costs, there are several factors that drive the total cost of ownership (TCO). These include significant and generic factors such as energy costs, floor space costs, infrastructure hardware costs, cluster management complexity, scalability, and people costs. Other factors such as costs for application porting and migration (in case of hybrid CPU-GPU servers), retraining of the IT staff, software licensing, unplanned outages and solution deployments are also typically incurred or considered. Newer considerations include 'Carbon tax' and the regulatory inevitability of reducing energy consumption[6] . Our analysis shows that RAS costs can be significant for Petaflop scale supercomputing clusters.

In this paper, we evaluate the air cooled iDataPlex dx360 M4 solution from IBM with other standard rack based x86 clustered servers with Intel Westmere processors for HPC from the perspective of Total Cost of Ownership (TCO) over a three-year period. In order to maintain relevance and objectivity across diverse industry scenarios, in this study we consider common factors such as: energy costs, floor space and data center costs, hardware acquisition costs, downtime costs, and people costs.

The key platform differentiators for the iDataPlex dx360 M4 system are innovative power and cooling mechanisms, a half depth rack design that substantially reduces floor space and aisle related footprint, and the latest Intel Xeon E5-2600 (Sandy Bridge) processor series. Together, these significantly reduce site costs and help achieve higher energy efficiency, especially at Petaflops and beyond. dx360 M4 servers have unparalleled scalability and smaller footprint (high package density). However, relative to standardized commodity pure x86 Cluster architectures which are widely used today for HPC, the dx360 M4 requires higher initial capital commitment but brings significant savings in the form of reduced energy consumption and site infrastructure costs. In fact, our analysis in the few Petaflops regime, indicates that the overall TCO of dx360 M4 based systems is a little less than half of an equivalent x86 (Intel Westmere) based HPC cluster.

Today, the iDataPlex is being used at several leading supercomputer sites in Top500 list running many industrial applications ranging from financial services to the life sciences delivering unsurpassed performance, scalability, energy efficiency, and substantial reduction in data center footprint and costs.

On the path to Exascale, as the die casts shrink, exploding cores and costs increase the complexity to unprecedented levels. The table below depicts evolution of supercomputing clusters, the growth in the number of cores per node and the overall cluster peak performance in (peak flops) over the years.

| Typical Supercomputing Cluster configuration over the years | 2001 | 2005 | 2005 vs. 2001 | 2010 | 2010 vs. 2005 (multiplier) |
|---|---|---|---|---|---|
| Cores per node | 4 | 16 | 4 | 32 | 2 |
| Compute nodes | 640 | 567 (8S) | 0.89 | 3442 (4S) | 6.07 |
| Number of cores | 2560 | 9072 | 3.54 | 110,132 | 12.14 |
| Core performance | 1.95 GF | 6.61 GF | 3.39 | 9.08 GF | 1.37 |
| Node performance | 7.81 GF | 106 GF | 13.56 | 291 GF | 2.75 |
| FLOPs peak | 5 TFlops | 60 TFlops | 12 | 1 PFlops | 16.67 |

If you look at how electrical and space costs are evolving[2] , trends show that performance/price (FLOPS/$) is increasing faster than the facility needs in terms of FLOPS/sq. feet and FLOPS/W. With advances in cooling

[6] Preparing for green regulations: http://insidehpc.com/2009/07/16/big-datacenters-prepare-now-new-green-regulations/

technology, vendors are trying to control both the energy consumption by the supercomputers as well as the energy required to cool the systems given high compute density in today's supercomputing cluster. Newer techniques such as liquid cooling allow higher density and cooling efficiency with less floor space use but could result in higher oveall electricity consumption[7].

## Trends in High Performance Technical Computing

As Technical Computing continues to grow faster than the overall server industry, solution providers have geared to engineer breakthroughs in performance, scalability, price/performance, space and energy costs and software for better systems and applications management.

The anticipated energy costs to power and cool large HPC data centers are likely to increase more rapidly during the next decade unless economical approaches to energy production are developed in the near term. The IT industry is defining additional metrics such as gigaflops/watt, PUE[8] which rate the HPC solution providers and data center operators today. As HPC solution providers compete fiercely for bragging rights today, the Green 500 list is becoming as important as the Top 500 list of supercomputers.

In recent years, HPC solution providers have made significant innovation in "green" and next generation data centers to reduce the TCO, reflecting not just capital costs but operational and maintenance costs as well. More than an economic pain point or a social responsibility, bringing sound environmental principle to bear in operating the data center can become a competitive advantage and a source of operational stability and increased reliability. It is from this perspective that we compare prominent HPC supercomputing systems available today.

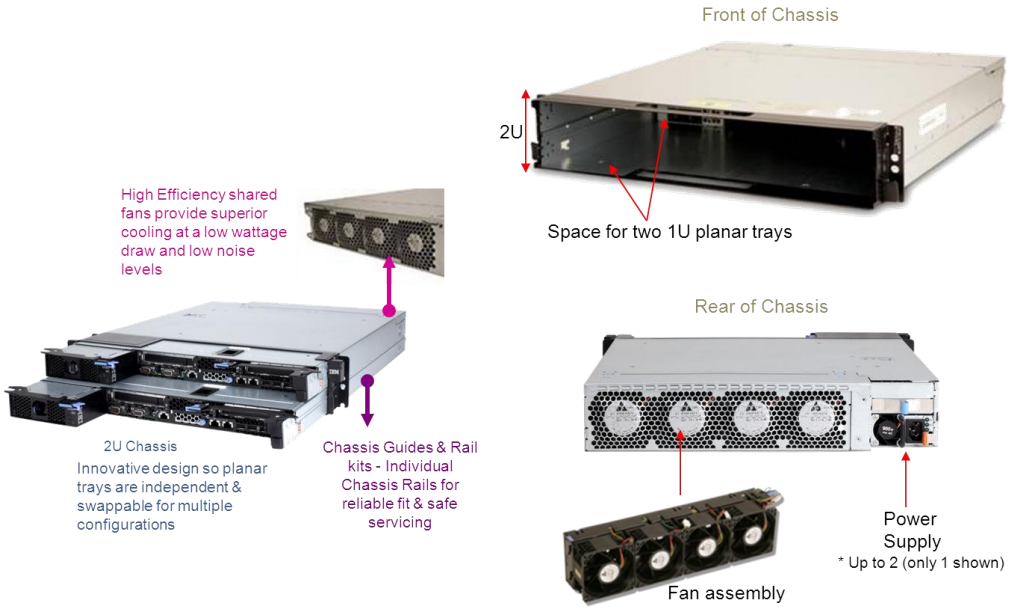


**Figure 3: The iDataPlex dx360 M4 Salient Features (Source: IBM)**

## iDataPlex dx360 M4 – Innovative and Energy Efficient

Next-generation scale-out data centers are challenged by power, cooling, floor space, acquisition and maintenance costs, and will be subject to various government regulations for power efficiency, carbon emissions, and environmental impact. To address these concerns and to design the next-generation data center model, IBM announced the Project Big Green initiative. IBM's System x iDataPlex dx360 M4 is one such technology, which defines a new server architecture based on the iDataPlex 100U rack cabinet, Flex Node and Chassis, Rear Door Heat exchanger, and a rack management appliance. The 100U rack cabinet is a standard-sided rack, turned 90 degrees and holds nearly 2.5x as much equipment as before, all with higher density than with standard rack servers, along with lower energy usage and cooling needs.

[7] Horst D. Simon: Petascale systems in the US http://acts.nersc.gov/events/Workshop2006/slides/Simon.pdf

[8] Power Usage Effectiveness (PUE) = Total facility Power / IT equipment power consumption http://en.wikipedia.org/wiki/Power_usage_effectiveness

## iDataPlex dx360 M4 Innovations

IBM has years of experience designing server technologies for both scale-up and scale-out settings that primarily focused on performance and scalability as the fundamental requirements. However, iDataPlex dx360 M4 also focuses on an additional set of goals:

- Significantly reduce the initial hardware acquisition costs and on-going maintenance costs for data center owners
- Dramatically improve efficiency in power consumption
- Virtually eliminate data center cooling requirements
- Achieve higher server density within the same footprint as the traditional rack layout
- Simplify manageability for massive scale-out environments
- Reduce the time to deployment through pre-configuration and full integration at manufacturing.

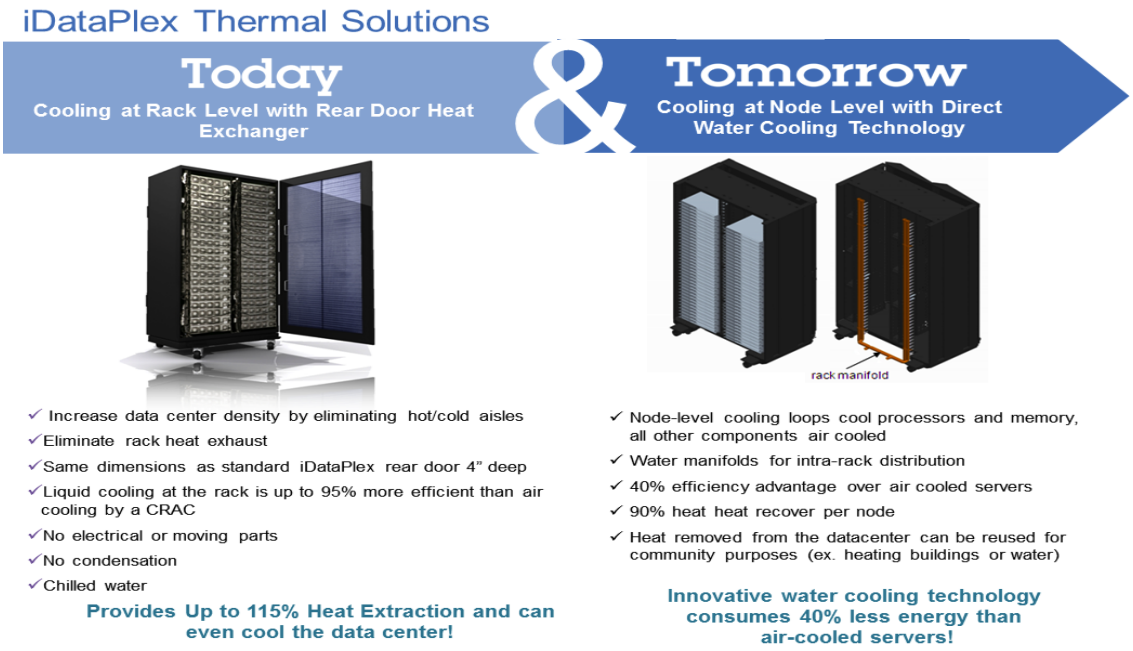


**Figure 3: iDataPlex dx360 M4 -Fast, cool, dense & flexible: TCO savings without compromise (Source IBM).**

In addition, dx360 M4 server technology addresses data center challenges at various levels with:

- A rack design that achieves higher node density within the traditional rack footprint
- *Flex node* chassis and server technology based on industry standard components
- Shared power and cooling components to improve efficiency at the node and rack level
- An optional Rear Door Heat eXchanger that virtually eliminates traditional cooling based on computer room air conditioning (CRAC) units
- Various networking, storage, and I/O options optimized for the rack design
- Intelligent, centralized management through a management appliance.

The iDataPlex rack is shallower in depth compared to a standard 42U server rack. The shallow depth of the rack and the dx360 M4 nodes are part of the reason that the cooling efficiency of dx360 M4 is higher than the traditional rack design - air travels a much shorter distance to cover the internals of the server compared with airflow in a traditional rack.

A single iDataPlex dx360 M4 100U rack enclosure provides 2x (or more) compute density of a standard 42U rack enclosure within the footprint of a single 42U rack. The iDataPlex rack has 84 horizontal 1U slots for server chassis and 16 vertical slots for network switches, PDUs, and other appliances. The rack is oriented so that servers fit in side

by side on the widest dimension. For ease of serviceability, all hard drive, planar, and I/O access is from the front of the rack. There is little need to access the rear of the iDataPlex rack for any serviceability other than to service the Rear Door Heat eXchanger. In addition to this, the iDataPlex 2U chassis supports various configurations, including 2 compute nodes, or 1 compute node and 1 storage node, or 1 compute node and 1 I/O node, or 1 compute node and 1 GPU node, and configurations can be changed as needed. That flexibility/interchangeability are major innovations of iDataPlex. It's sort of a hybrid of typical rack servers and blade servers.

Key development objectives for iDataPlex are to lower acquisition and operating costs and simplify the process for customers looking to build large-scale data centers.

In the next section, we present a detailed TCO analysis of Petascale Technical Computing clusters comparing two types of cluster architectures – one, typical x86 rack servers (using Intel Westmere processors) and two, the latest IBM iDataPlex dx360 M4 (using Intel Xeon E5-2600 processors series – Sandy Bridge). The iDataPlex is much more energy efficient. Despite IT capital cost advantages of x86 commodity servers, the Site Infrastructure costs are significantly higher as we scale typical x86-based servers to Petascale. In fact, our study indicates that for a 2.954 Petaflop range cluster using iDataPlex (Intel Xeon E5-2600) processor, the overall TCO is 57% lower than a comparative cluster using typical x86-based servers with Westmere processors.

## The TCO Methodology

For this study, we created a TCO model for cluster based supercomputing systems using the Uptime Institute's data center TCO calculator as the base. It was further enhanced for RAS metrics based on an earlier study by Cabot Partners on the Blue Gene and RAS aspects of HPC clusters[9]. A set of anchor systems were identified from several systems listed in the Top500 and Green500 lists and public information spanning a range of 30 Teraflops to 5 Petaflops of sustained performance and covering both architectures under study – typical x86 clusters and IBM iDataPlex dx360 M4 systems. For each of these anchor systems, data related to their configuration and performance metrics were fed into our TCO model and subsequent results obtained were analyzed.

**Anchor Systems and Scaling Assumptions**

In this study, we chose the following supercomputing systems as the anchor systems. Our criteria included those systems which are listed in Top500, Green500 lists and which ranged from tens of Teraflops to 5 Petaflops in performance. Within this range, we shortlisted systems which cover both kinds of architectures – typical x86 (Intel Westmere processors), and iDataPlex dx360 M4 using Intel E5-2600 series processors. We did not include systems with Hybrid CPU-GPU boards so clusters such as Pleiades, Tianhe 1A are excluded. This study is focused on non-hybrid systems that do not require extra code migration, porting and re-training efforts. Our TCO calculator required several configuration details for these supercomputers. We selected these systems as their configuration information and other details required for TCO were publicly available online for a candid analysis.

The table below shows a list of anchor systems and their performance rating (FLOPs). We scaled down the iDataPlex dx360 M4 Petaflop system to the x86 anchor system performance levels, 35.8TF and determined the TCO for each for a fair comparison at Teraflop scale. Also, we scaled up the x86 based cluster to Petaflop scale, 2.954 PF which matches the dx360 M4 anchor performance and calculated the TCO for each for a fair comparison of both architectures at Petaflop scale. This assumption is actually very favorable (from a TCO perspective and an acquisition cost perspective) towards x86 clusters and makes our assumptions very conservative and more favorably biased towards x86 clusters.

| | **Teraflop Clusters** | **Petaflop Clusters** |
|---|---|---|
| **iDataPlex clusters (Intel Xeon E5-2600 series processors – aka Sandy Bridge)** | | SuperMUC[10], iDataPlex dx360 M4 ( 2.954 PF |
| **x86 clusters (Intel Xeon 5600 series processors aka Westmere EP)** | Red Sky[11], Sandia National Labs, 433.5TF | |

[9] Cabot RAS Study: http://www-03.ibm.com/systems/resources/systems_deepcomputing_IBMPower-HPC-RAS_Final-1.pdf

[10] SuperMUC is located in the Leibniz Supercomputing Centre in Garching, Germany. http://www.lrz.de/services/compute/supermuc/systemdescription/

[11] Red Sky is an x86 based supercomputer location at Sandia National Labs, US. http://i.top500.org/site/3121

## Key Findings – Factors that Fuel TCO at Petaflop Scale and Beyond

The TCO consists of several significant factors such as electricity which accounts for the annual energy cost of wattage being consumed, floor space costs that accounts for the density, architectural costs associated with infrastructure including networks and storage, cabling, capital hardware acquisition cost, and the people costs were restricted only to system maintenance at the customer site in the data center. Here we do not consider factors like application enablement and migration costs, operating costs and training costs for all systems as these costs are similar for the systems investigated. We also did not consider the costs involved with upgrading equipment but scalability and reliability were taken into account. Software licensing costs vary across providers and industries and hence were not considered.

**Energy Consumption:** With technological and performance advancements in processor technology and design that enable high-density packaging of processing cores in a server, processor energy consumption is on the rise. A larger number of transistors translates into increased operational power costs and higher levels of heat generated per chip. The iDataPlex servers demonstrated a cost advantage of approximately 80% reduction in energy as compared to typical x86-based cluster.

**RAS:** As the temperature rises, the system failure rates increase. In addition to temperature, as the data center footprint grows with the increase in number of sockets, cores, nodes etc, the system failure rates rise even more. Hence adequate cooling and RAS management is essential for efficient functioning of larger data centers. In addition to the direct cooling costs involved, cooling systems could occupy additional space on racks. Each rack cannot be fully populated with only server nodes, and more racks would be needed for a particular performance level. As the energy requirements in large data centers rise, additional UPS and backup power capacities are needed for the operation and cooling of the data centre.

**Total Floor Area:** In the last 5 years, the performance of systems in data centers has increased exponentially. Advanced networking technologies and high speed InfiniBand switches have enabled clustering of a large number of nodes. Most equipment layouts are in a single row of rack-mounted servers, forming aisles between them. Network switches and storage devices, placed alongside the racks, are often as big as the racks themselves. This has caused a significant strain on the infrastructure of data centers that were built for hardware with much less capability than what is being shipped today. The much higher rack power levels have caused customers to spread out their server products in order to cool them in the current facilities, using up valuable and expensive raised floor space.

The electrically active floor area of a data center is estimated to be only about 40% of the total floor area of the data center. Chillers, fans, pumps, service aisles between racks and other electrically inactive components make up the remaining space in a data center.

**IT acquisition costs:** The IT-related capital cost, is Capital Recovery Factor times the capital costs incurrent for a 3 year life. It accounts for the total IT-related capital cost incurred from investment in total number of filled racks, internal routers and switches, and rack management hardware. At a low level, the total IT-related capital cost for the investment in total number of filled racks includes the acquisition cost for servers, disk and tape storage, and networking.

**Other Facilities Costs:** Other facilities costs include interest during construction estimated based on total infrastructure and other facility capital costs at a fixed rate of interest, land costs, architectural and engineering fees, and inert gas fire suppression. Land costs are based at $100,000 per acre and architectural and engineering fees are estimated at 5% of kW related infrastructure costs plus other facility costs (electrically active).

**Operating Expenses:** The IT and site-related operating expenses account for the total operating expenses incurred in investing in each of the two systems under study – iDataPlex dx360 M4 and x86 (Westmere) based clusters. Operating expenses demonstrated a marked rise for typical x86-based Petaflop clusters.

**Total costs:** The summary of both supercomputing cluster architectures at the performance levels undertaken in this study finds the dx360 M4 to be the most cost-effective system of choice which is significantly less than the typical x86-based CPU only clusters by more than 50%.

**Other Considerations:** Typical x86-based cluster architecture is the most prevalent offering in the HPC market with a thriving ecosystem of applications, software tools, and other HPC components. The same is not true for hybrid CPU-GPU based clusters due to complexity involved with CUDA programming models and migration costs involved. The costs of application migration, training, and deployment for this architecture could be significantly less for CPU only

(no GPUs) clusters but the scalability, costs and reliability issues could easily surpass those.

Energy costs and cooling costs have been a major buzz in the IT industry but our study indicates that energy costs are a small component (<10%) of the overall Total Costs of Ownership of a supercomputing cluster. However, energy considerations could be a limiting factor for the growth of HPC data center capability in urban and semi-urban locations. This is validated by trends whereby large computing clusters for supercomputing or cloud computing are being located close to sources of power across the globe like Google's data center or Microsoft's new data center or those located near hydro-electric power plants in China. Floor space restriction in facilities coupled with power requirements make the dx360 M4 a very attractive choice in the HPC data centers. Next we examine the TCO for the anchor systems from Teraflop to Petaflop range of performance levels.

## TCO Analysis

IBM's iDataPlex dx360 M4 systems powered by Intel Xeon E5-2600 processors (Sandy Bridge) are much more energy efficient and have 57% lower overall TCO for Petaflop scale supercomputing clusters as compared to x86-based clusters. The IT Capital costs are higher for the dx360 M4 but the Site Infrastructure and Energy costs taken together for typical x86-based rack servers using Intel Westmere chips are significantly higher as shown in the following set of comparative TCO charts.

First, let's examine the Price/Performance for the anchor clusters in Figure 5. The SuperMUC cluster using iDataPlex dx360 M4 servers with Intel Sandy Bridge processors is the clear leader in the Petaflop regime.

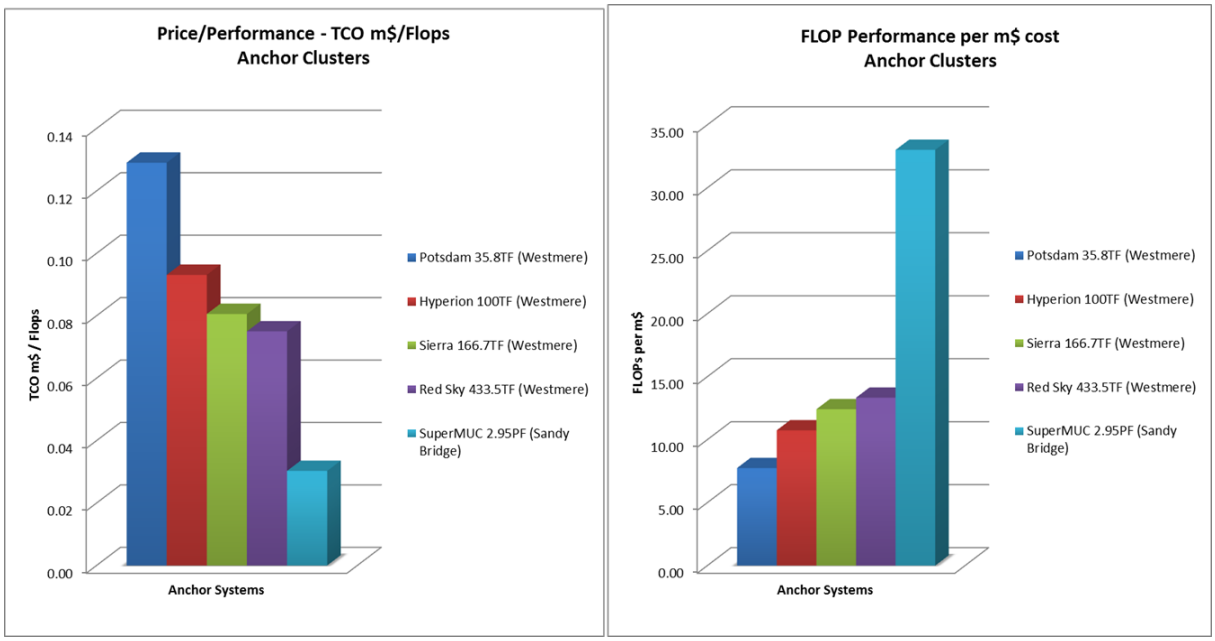


**Figure 5: Price-Performance of Anchor Clusters**

Next we present – Figure 6 - the comprehensive summary of TCO data including the individual TCO components for all the anchor systems used in this study.

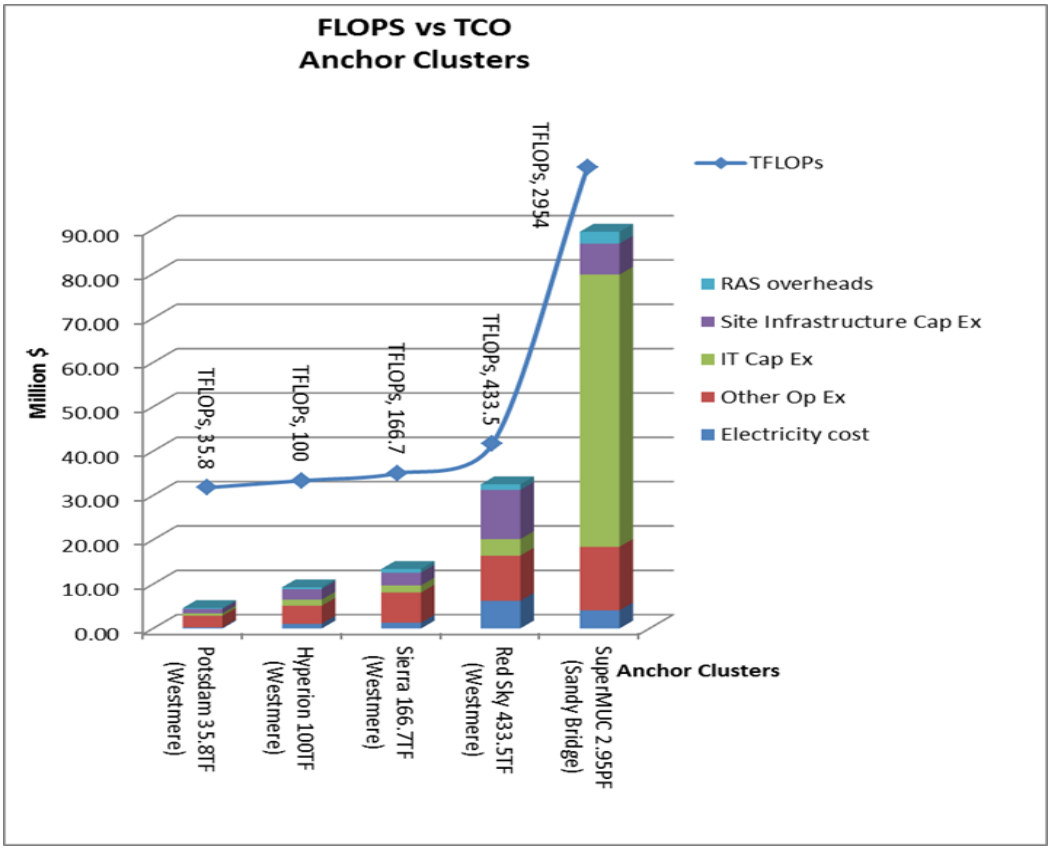


**Figure 6: Anchor Systems Comparative overall TCO**

We then detail the TCO data for anchor systems in the following set of pie charts. From the anchor systems pie chart (Figure 7), it is evident that for typical x86-based rack server clusters at Teraflops scale, Other Op Ex costs constitute a major component of TCO whereas at for the dx360 M4 Petascale systems, IT Capital costs form the major component of TCO.

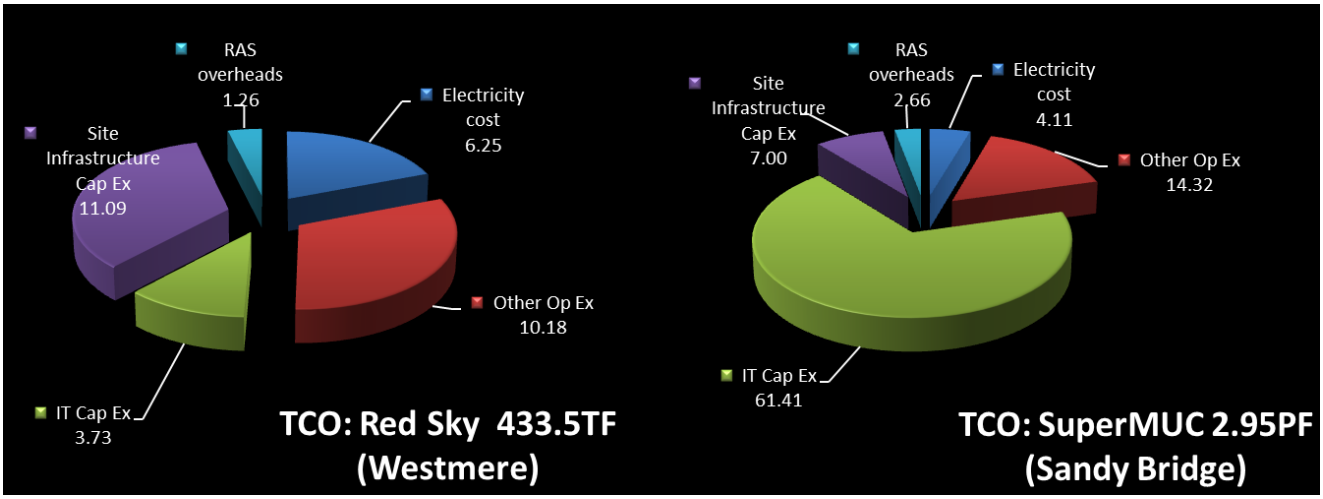


**Figure 7: Anchor Systems Comparative overall TCO Pie Charts**

The following charts and figures indicate that at the higher Petaflop range, energy costs shoot up significantly for Westmere based x86 commodity cluster systems. The dx360 M4 clusters are the most energy efficient and have the lowest RAS overheads – greatly lowering the overall TCO for these systems.

## FLOPS vs. TCO

The overall TCO for x86 cluster system rises almost exponentially from Terascale to Petascale. IBM's dx360 M4 has a lower overall TCO especially at Petascale as shown in following charts.

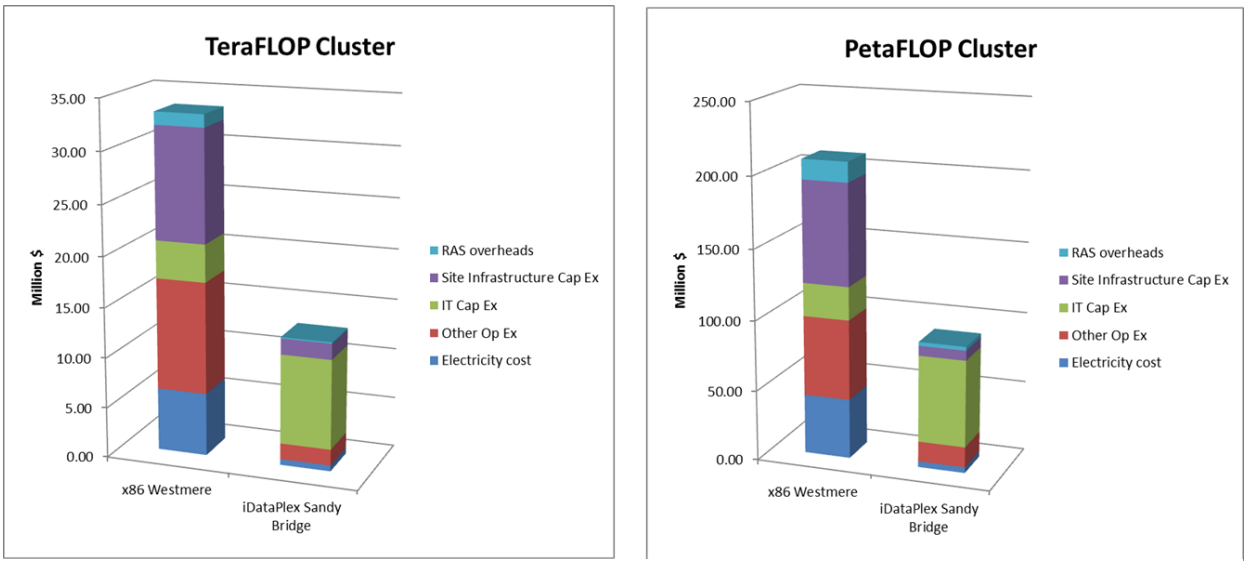


**Figure 8: Comparative TCO charts**

The following pie charts show the TCO vs. Performance for different architectures for Teraflop as well as Petaflop scale. Below the 100TF scale, the TCO of typical x86-based (only CPU, no GPU) clusters is comparable to iDataPlex clusters. But when scaling up to higher performance systems i.e. Petaflops, the Energy and Operating Expenses rise up significantly for x86 clusters; making the dx360 M4 a winner in terms of price/performance, MFLOPS/W and MFLOPS/W/sq. feet of data center floor space. At these higher performance levels, the dx360 M4 based cluster's total costs are much lower reflecting its fundamental advantages of an ultra-scalable architecture, leading-edge technology design and energy optimization for supercomputing needs.

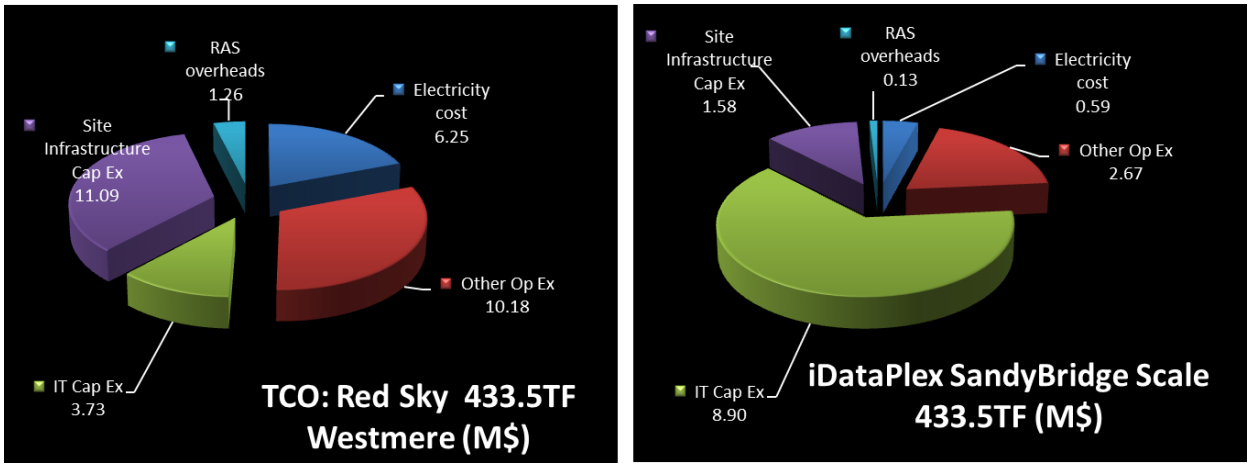


**Figure 9: Comparative TCO at Terascale**

At Terascale, both x86 clusters as well as the dx360 M4 based systems have similar individual TCO components. The iDataPlex is much more energy efficient than x86-based clusters. Individual TCO components such as IT Cap Ex, Site Infrastructure costs, Other Op Ex are similar. dx360 M4 cluster is marginally better than x86 clusters in RAS costs.

At Petascale, Site Infrastructure costs are the major component of TCO for typical x86-based clusters whereas for iDataPlex, the IT Capital costs are the major component of overall TCO, as shown in the following pie charts.

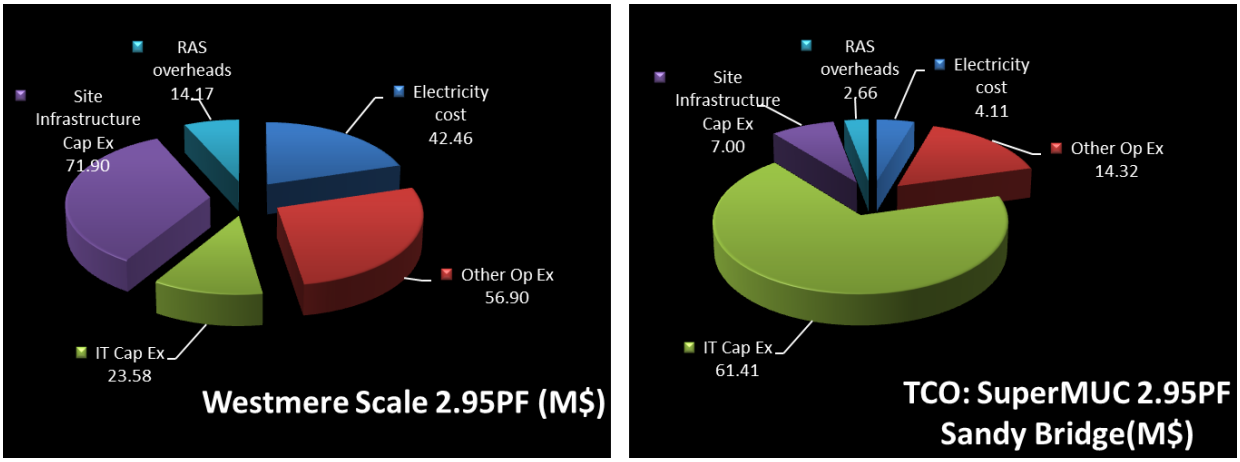


**Figure 10: Comparative TCO at Petascale**

Next, we compare the Energy costs for different systems studied as part of this TCO analysis.

## Energy Costs

It is clearly evident that as the number of cores, sockets and nodes increases for Petaflop systems, the Energy requirements rise significantly. As compared to typical x86-based commodity clusters, the iDataPlex leads in energy efficiency and the overall TCO both at Teraflop scale as well as at Petaflop scale of clusters.

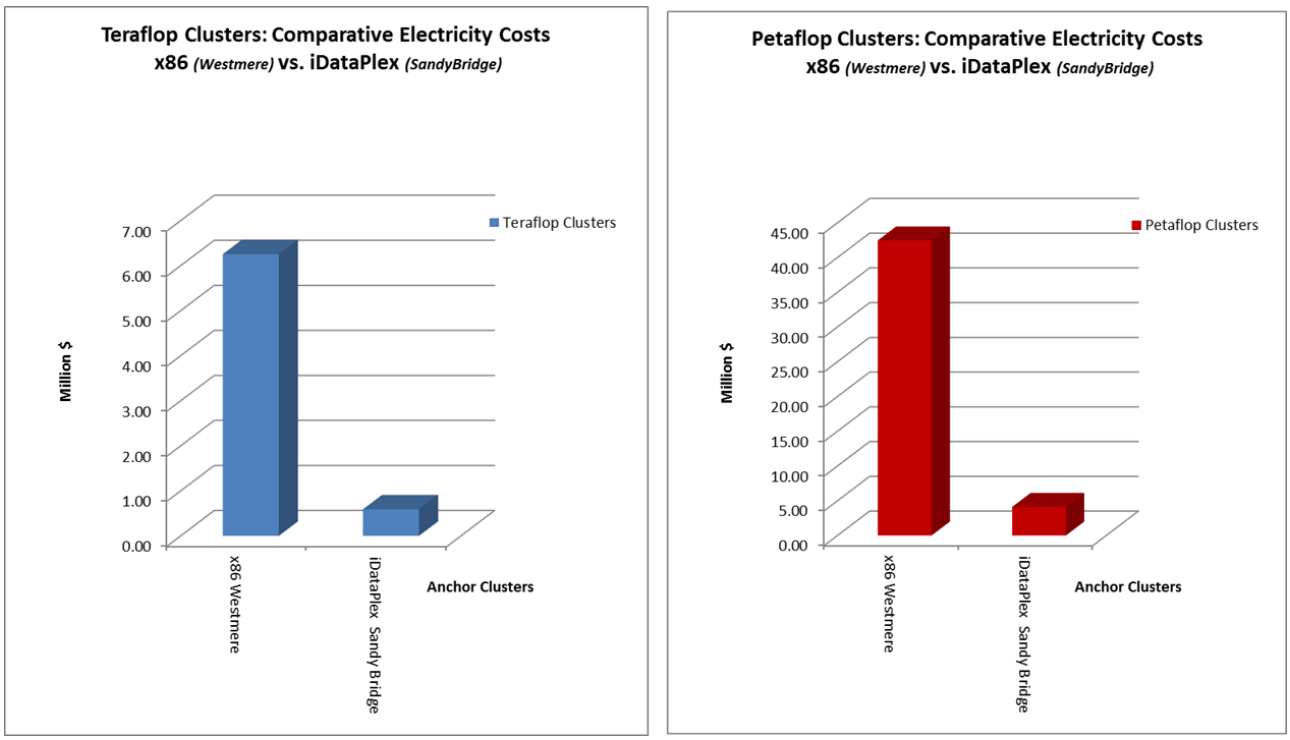


**Figure 11: Comparative Electricity Costs - x86 (Westmere) vs. iDataPlex (Sandy Bridge)**

At Terascale, Energy Costs constitute 6% of the overall TCO for x86 clusters whereas for the iDataPlex dx360 M4 it is 2% of overall TCO. At Petascale, Energy Costs constitute 20% of overall TCO for typical x86-based HPC clusters whereas for dx360 M4 it is 4% of overall TCO. What this means is that if you scale a cluster from Teraflops to Petaflops, x86 commodity server based clusters would see more than three times rise in Energy costs whereas the iDataPlex would see approximately a doubling of Energy costs with the overall energy costs still a significantly smaller component of TCO as compared to typical x86-based rack server clusters.

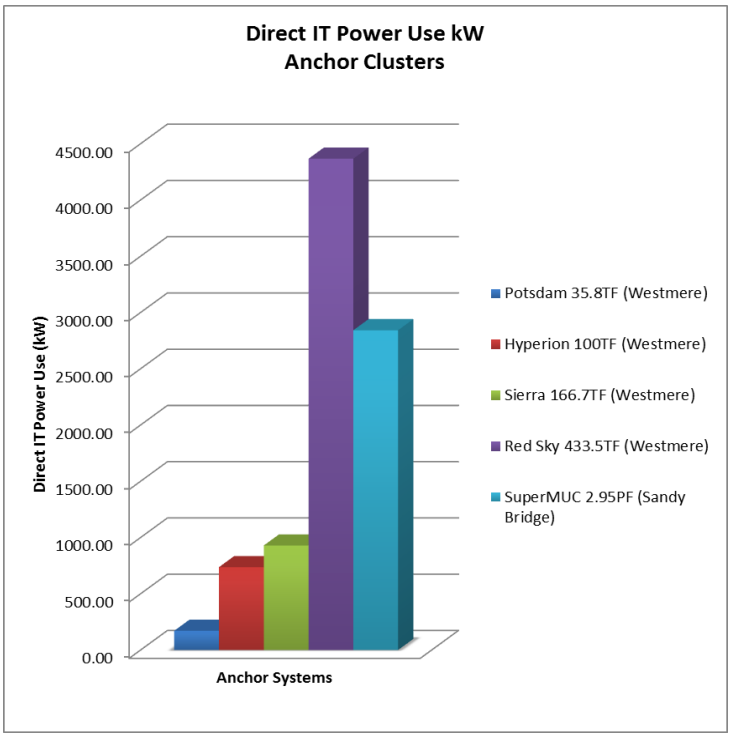


**Figure 12: Direct IT Power Use (kW) in Anchor Clusters**

## Infrastructure Costs vs. IT Costs

If we look at the Infrastructure vs. IT costs as a percentage of TCO, in Teraflop scale clusters, IT related costs are lower than Infrastructure related costs for both typical x86-based server clusters and iDataPlex systems. But at Petascale, IT related costs constitute 80% of overall TCO for iDataPlex dx360 M4 server (using Intel E52600 processor series - Sandy Bridge) clusters whereas it is 35% of overall TCO for x86 (Westmere) based clusters.

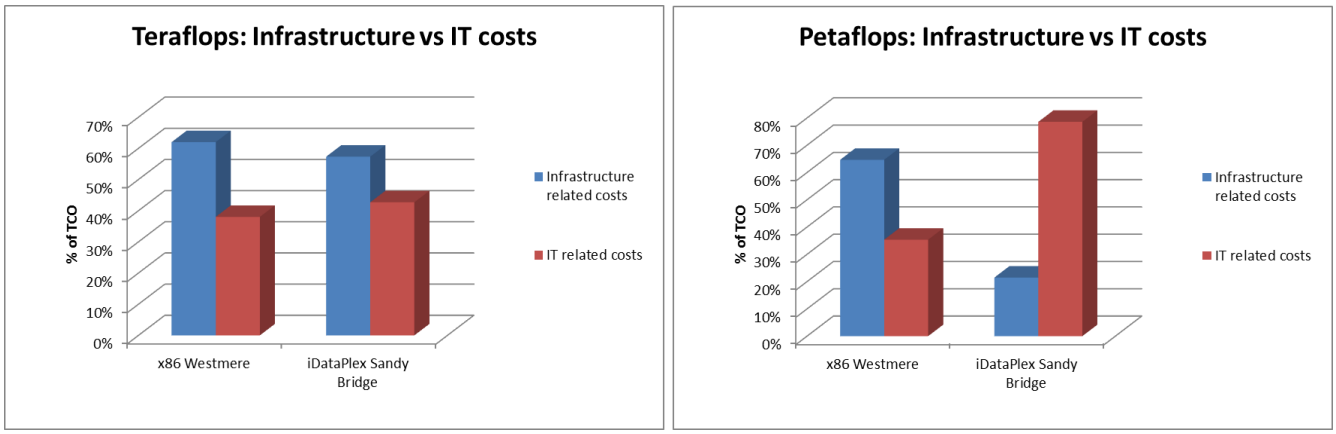


**Figure 13: Infrastructure vs. IT costs Comparison - x86 clusters (Westmere) vs. iDataPlex (Sandy Bridge)**

## Results and Concluding Remarks

This TCO study includes the energy costs for the entire data centre, the IT capital costs and infrastructure capital costs annualized over a 3 year life and a 15 year life for land and other fixed property, and the annualized operating

expenses. We studied system configurations including the floor space needs, power usage, performance, number of racks, cores, and power used per rack at 30TF, and 2.954 Petaflops. We looked at real systems (referred to as anchor systems) and used the TCO calculator to arrive at the various TCO cost estimates. Based on our study, IBM's System x iDataPlex dx360 M4 servers powered by Intel Xeon E5-2600 processor series (Sandy Bridge) are much more energy efficient and cost effective for Petascale HPC clusters as compared to typical x86-based servers using Intel Westmere processors.

The dx360 M4 systems are attractive because of the total performance it offers, and its total cost of ownership is far lower than the other systems in the study, over the three-year period. Even then, the appetite for HPC is growing by a factor of 10 (from 10 cores in 1992, 100 cores in 1998, 1000 cores in 2004, 10,000 cores in 2010 and a realistic projection of 1.5 million cores by 2018), hence the biggest challenge is to make the systems more efficient, scalable, and reliable from the perspective of energy, floor space, operating expense and deployment costs. iDataPlex has the appropriate architecture to match and exceed these future requirements. Furthermore, the lower energy consumption, innovative power and cooling features and smaller footprints of iDataPlex systems especially at larger performance levels could significantly increase system reliability and hence reduce downtime costs.

Commodity x86 clusters cost less to buy. However, at Petaflop scale there are other costs which influence the TCO by a much higher value for x86-based server clusters; tipping the scale in favor of the iDataPlex dx360 M4. The Energy Costs and Operating Expenses for x86-based clusters are extremely high - dx360 M4 has 57% lower overall TCO, 45% lower Op Ex and over 80% lower energy costs compared to an equivalent commodity x86 cluster in the Petaflops range. For Technical Computing environments that require large scalability/performance and economical operation, iDataPlex is an excellent platform. The higher capital acquisition costs at Petascale as compared to x86 commodity clusters are insignificant compared to the sharp rise in energy and operating costs of commodity clusters.

Over the last decade, with the widespread penetration of industry-standard clusters, HPC capital expenses as a percentage of IT spend have decreased. But the associated operational expenses to manage these higher computing density HPC data centers have escalated largely because of increased costs in systems administration, managing RAS, energy, facilities and cooling as these systems have become denser with thousands of cores. Also, memory density has increased greatly as well with technology evolution. The DIMMs use a lot of energy. The past few years have seen a progression from 1.8V DDR2 memory to 1.5V DDR3 memory to 1.35V DDR3. Operating applications using hundreds or thousands of terabytes of memory requires a lot of energy, but systems with 1.35V memory, such as iDataPlex, consume up to 19% less energy[12] per DIMM than those with 1.5V memory. In addition to the leading-edge Intel Xeon E52600 processor series (Sandy Bridge) powered iDataPlex dx360 M4 server, the rear door heat exchanger feature and the latest innovative direct water-cooling in these servers are designed to further lower data center TCO. This will help rein in these escalating operational expenses while managing capital expenses and protecting a customer's investment in applications, skills, and programming models. Overall, IBM's iDataPlex dx360 M4 servers help HPC supercomputing clusters achieve better energy efficiencies and lower TCOs for the Exascale computing era.

## For More Information

The cost drivers of TCO are quantified based on the model provided by the Uptime Institute, Inc.: http://uptimeinstitute.org/content/view/57/81

**Assumptions**

1) % of rack filled based on Uptime consulting experience data is available at: http://www.missioncriticalmagazine.com/MC/Home/Files/PDFs/(TUI3011B)SimpleModelDeterminingTrueTCO.pdf .
2) Energy use per U taken from press releases, public presentations and other online public info. Server power and costs per watt assumes IBM iDataPlex system.
3) Energy use per rack is the product of the total number of Us filled times watts per installed U.
4) Total direct IT energy use is the product of watts per rack times the number of racks of a given type.
5) Cooling electricity use (including chillers, fans, pumps, CRAC units) is estimated as 0.65 times the IT load.
6) Auxiliary's electricity use (including UPS/PDU losses, lights, and other losses) is estimated as 0.35 times IT load.
7) Total electricity use is the sum of IT, cooling, and auxiliaries. Cooling and auxiliaries together are equal to the IT load (Power overhead multiplier = 2.0).
8) Electricity intensity is calculated by dividing the power associated with a particular component (e.g. IT load) by the total electrically active area of the facility.
9) Total electricity consumption is calculated using the total power, a power load factor of 95%, and 8766 hours/year (average over leap and non-leap years).
10) Total energy cost calculated by multiplying electricity consumption by the average U.S. industrial electricity price in 2011 as per http://www.eia.doe.gov/oiaf/aeo/pdf/0383 (2010).pdf (8.6 cents/kWh, 2008 dollars).
11) Watts per thousand 2011 dollars of IT costs taken from selective review of market and technology data. Server number calculated assuming IBM iDataPlex public information available online.
12) External hardwired connections costs are Uptime estimates.

[12] 1.35V memory used in iDataPlex uses 19% less energy: Source (IBM)

13) Internal routers and switch costs are Uptime estimates.
14) Rack management hardware costs are Uptime estimates.
15) Total costs for racks, hardwired connections, and internal routers and switches are the product of the cost per rack and the number of racks.
16) Cabling costs totals are Uptime estimates.
17) Point of presence costs are Uptime estimates for a dual POP OC96 installation.
18) kW related infrastructure costs (taken from Turner and Seader 2006) are based on Tier 3 architecture, $23,801 per kW cost. Assumes immediate full build out. Includes costs for non-electrically active area. Construction costs escalated to 2009$ using Turner construction cost indices for 2010 and 2011 (http://www.turnerconstruction.com/corporate/content.asp?d=20) and 2011forecast (http://www.turnerconstruction.com/corporate/content.asp?d=5952). Electricity prices escalated to 2009$ using the GDP deflator 2009 to 2010 and 3% inflation for 2010 to 2011.
19) RAS costs are based on inputs whereby hourly cost of downtime ranges from thousands to millions of dollars across applications, industries and companies. We are taking a conservative number of $1000 - and multiplying it by total downtime for the system under investigation. It was further scaled depending upon the total people time involved in the cluster IT maintenance and up-keep.

**IBM iDataPlex Data Sources**

1. IBM Sales presentations
2. SUPERMUC http://www.theregister.co.uk/2011/01/03/prace_lrz_supermuc_super/, http://www.lrr.in.tum.de/~gerndt/home/Teaching/Parallel%20Programming/superMUC.pdf , http://www.lrr.in.tum.de/~gerndt/home/Teaching/Parallel%20Programming/superMUC.pdf, http://www.nm.ifi.lmu.de/~kranzlm/vortraege/2011-10-17%20Linz%20-%20SuperMUC%20-%20PetaScale%20HPC%20at%20the%20Leibniz%20Supercomputing%20Centre.pdf , http://www.lrz.de/services/termine/it-betrieb/gottschalk_2011-12-01.pdf , http://www.lrz.de/services/compute/supermuc/systemdescription/
3. Yellowstone http://www.unidata.ucar.edu/community/seminars/Yellowstone_Briefing_UCP_20111129.pdf , https://www2.cisl.ucar.edu/book/export/html/1228

**x86 Cluster Data Sources**

1. http://www.extremetech.com/article2/0,2845,2362102,00.asp
2. http://www.openfabrics.org/archives/spring2010sonoma/Wednesday/9.00%20Marcus%20Epperson%20Red%20Sky/OFED%202010%20RedSky%20IB.pdf
3. http://www.hpcwire.com/features/Intel-Ups-Performance-Ante-with-Westmere-Server-Chips-87822922.html?page=2
4. http://www.top500.org/system/performance/10584
5. https://newsline.llnl.gov/_rev02/articles/2010/sep/09.24.10-sierra.php
6. http://www.theregister.co.uk/2011/01/03/prace_lrz_supermuc_super/print.html
7. http://publib.boulder.ibm.com/infocenter/idataplx/documentation/topic/com.ibm.idataplex.doc/dg1bdmst.pdf