August, 2008

Achieving Superior Performance per Watt With SuperBlade® and 1U Twin™ Server Systems

Supermicro Delivers the High Performance and Power Savings Data Centers Need to Stay Ahead of the Game

Contents

Executive Summary

A significant transformation is underway in the information technology industry. Enterprises are utilizing their server systems more effectively to improve data center performance and to strengthen their green footprints. Servers offering both high performance and power efficiency are key technologies allowing data center managers to deliver more information to their users with improved TCO and environmental impact.

These changes are essential to control rising data center costs and to improve the accuracy and availability of information. In addition core IT systems and applications must be highly scalable to support rapidly growing workloads, resilient to ensure critical information is always available to those who require it, and flexible enough to adapt quickly as requirements grow. They must also be cost-effective and easily managed, so that the advantages of increased operational efficiency are realized.

Blade servers and twin server systems are ideal for these demanding requirements. They offer the high-end scalability and availability of traditional server systems, but without the higher cost and power requirements. As computing demands continue to grow, Blade and twin-based systems such as the Supermicro SuperBlade® and 1U Twin™ will be increasingly valuable assets for scaling and adapting core applications, managing rapidly expanding data volumes and satisfying electric power and environmental challenges.

Economic Impacts of Performance per Watt on TCO

The subject of 'Performance per Watt' continues to gain importance for the management of data centers. This is not surprising given rapidly escalating electricity costs, exploding end user demand for IT resources, electrical conduit exhaust in many metro areas, and environmental concerns. The search for the ultimate figure of merit has focused on various efficiency measures, most recently Performance per Watt. This metric has gained acceptance because of its clear effects on data center TCO (Total Cost of Ownership) involving direct operating expenses, capital costs, and opportunity costs.

Direct Operating Expenses – The most direct impact of the cost of electricity is on operating expenses associated with IT equipment power and cooling. It may be useful to set the scale with a simple example. One watt of power consumed continuously over the course of a year at $0.075 per kW-Hr [1] with an 83% overhead for cooling and auxiliary equipment [2] [3] [5] costs about $1.20 per year.

1 Watt-Year × 8760 Hr / Year × 0.075 $ / kW-Hr × 1.83 = $1.20

Assuming a 20 to 25 Watt savings per server and 1,000 or more servers per data center, this direct impact becomes economically compelling. Thus server systems that consume less power and require fewer cooling resources are often considered more valuable to the enterprise. Further, systems that operate at lower temperatures are more reliable (the failure rate of semiconductor components typically doubles for every 10 to 15 °C increase in operating temperature).

Capital Costs – A far larger impact of 'Performance per Watt' comes when considering capital costs associated with data centers. Since data centers are operated within a business needs context they have expected lifetimes, predecessors and successors. The Uptime Institute has estimated that 45% of the cost of a new data center is for power and cooling equipment [4] [6]. Improved performance per watt allows a twofold impact with respect to capital expenses. First it allows the continued use of an existing data center that is already at its power and cooling limits for a few more years. Second it lowers the cost of power and cooling equipment per unit of computing when a new data center is built.

Opportunity Costs – Indirect impacts of improved performance per watt can be identified for data centers that are already at their power and cooling limits. For cases where business revenue is directly proportional to computing capacity, the ability to replace older servers with higher efficiency systems can have a large impact. A relatively small added TCO expense, in the form of higher efficiency servers, can drive outsized returns in revenue and operating profits.

Measuring Performance per Watt

Meaningful evaluations of performance per watt must begin with the workload run on the server and its relevance to actual business usage patterns. Four types of workloads are run to exercise the server systems in measuring performance per watt. The first is from the industry standard benchmark, SPECPower, by the SPEC organization, and is most applicable to Java based middle tier business logic. The second is the very well known SPECjbb2005, also by the SPEC organization and applies to server side Java business workloads. The third and fourth are the well known Linpack and Stream workloads that are most applicable to HPC data centers.

Linpack and Stream – Both Ends of the HPC Spectrum

The Linpack and Stream workloads provide complementary coverage of both ends of an HPC workload where the spectrum is parameterized by the ratio of 'core compute' to 'main memory access' time. Linpack is almost 100% CPU 'core compute' dominated while Stream is 90% or more 'main memory access' dominated. Measuring the power consumption of these two workloads gives a quick characterization of the two ends of the workload spectrum and provides estimates for upper & lower bounds of system power consumption.

Both of these workloads are well known throughout the HPC community and are frequently used in system procurement benchmarks. A good introductory description of the Linpack benchmark can be found at http://en.wikipedia.org/wiki/LINPACK and the references there. Details of the specific configuration used here can be found below. The web site for the Stream benchmark, http://www.cs.virginia.edu/stream/, contains a variety of information including source code, FAQ, results and commentary.

Optimized Performance per Watt

Supermicro has launched a wide range of server products based on the Intel® 5100 Memory Controller Hub Chipset. This chipset offers a number of features that make it an ideal candidate for building power optimized systems, including support for DDR2 memory and a simplified Southbridge. The Supermicro product lines leveraging this chipset include the new SuperBlade® blade server system populated with SBI-7125C-S3 and SBI-7425C-S3 blade modules, and the high density 1U Twin™ 6015TC-10GB dual server systems.

These systems achieve their world class Performance per Watt by leveraging advanced technology and system design expertise. High efficiency power supplies (PSUs) are designed to operate at significantly higher efficiencies of over 90%, well matched to the workload, greatly reducing energy losses. Motherboards are designed with leading-edge technology and high-end components such as high-efficiency VRMs (voltage regulator modules) to reduce energy consumption. Cooling subsystems including advanced technology heat sinks, pulse-width modulated fan speed controls, structured chassis airflow design, intelligent temperature and power management, are architected to most effectively cool all system components. Finally, efficient silicon building blocks, such as the 5100 chipset and low voltage processors, make large contributions to overall 'performance per watt'.

SuperBlade®: 296 GFLOPS/kilowatt Linpack Performance

Supermicro recently reported an industry-leading 290 GFLOPS / kW on the SuperBlade®: http://www.supermicro.com/newsroom/pressreleases/2008/press061708.cfm

Here we report the details of that testing as well as related testing of the 1U Twin™. The SuperBlade® was configured as follows:

- Enclosure – SBE-714D-R42, a 14 blade enclosure with 4x1400Watt PSUs. The PSUs have a particularly high 93% peak efficiency.
- Blades – The enclosure was equipped with 14 compute blades, each model SBI-7425C-S3. These are based on the 5100 chipset and were populated with 4x2GB DDR2-667 ECC, 2x Xeon L5420, 1x73GB SAS drive.
- CMM – The enclosure was equipped with a single Chassis Management Module, SBM-CMM-001.
- GbE – The enclosure was equipped with a single GbE switch, SBM-GEM-001.

A Chroma Programmable AC Voltage Source Model 61604 and Chroma Digital Power Meter Model 66202 were used for the power measurements. All measurements were made at 208V AC (rms) single phase. The final tests were run with 3x1400W PSU powered up. The fan speed control was set to 'Level 1' via the CMM interface. This provided adequate cooling (e.g. no processor throttling) with an inlet temperature of 25º C. and an outlet temperature of 40º C. The maximum power during the run is reported below.

Each system had RHEL5 installed. The Linpack binary used was the 64 bit binary from Linpack 10.0.2 available from the Intel web site as part of the MKL package. A script was written to remotely launch a single Linpack run on each blade via ssh. The Linpack binary then used OpenMP threads to parallelize across the cores on each blade. This procedure should not be confused with MP_Linpack, which runs in tightly coupled parallel mode across the blades with significant communication among them during the run. The test run here corresponds to an 'embarrassingly parallel' or 'throughput' test. This remains one of the most common usage models for commercial HPC today. The input parameters to the Linpack binary were: problem size was N=31,000, array leading dimension was 31,000, and the array alignment was 4KB.

The above configuration yielded an average 65.65 GFLOPS per blade, for 919.1 total GFLOPS across all 14 blades. The maximum power consumption was recorded as 3100 Watts AC. **This yields 296 GFLOPS / kW for the SuperBlade®.** The previously reported value of 290 GFLOPS / kW was thus conservative.

A highly optimized Stream binary run on the same configuration yielded 6.0 GBytes / Sec on the Triad portion for a total of 84 GBytes / Sec across all 14 blades. Power consumption was measured at a max of 2175 Watts AC, or 155 Watts per blade. Idle power consumption was 1365 Watts AC for the whole unit or 98 Watts per blade.

1U Twin™: 252 GFLOPS/kilowatt Linpack Performance

The HPC market segment is currently dominated by 1U server systems. The Supermicro 1U Twin™ has proven a popular choice for large HPC deployments due to its cost and power optimization. It is therefore instructive and appropriate to provide comparative measurements for that unit. The unit tested was the 6015TC-10G. This unit was equipped with a high efficiency 780W PSU and two X7DCT-10G server boards. Each unit was populated with 2x Xeon L5420 CPUs, 1 x 500 GB 7200 RPM SATA drive and 6x2GB of DDR2-667 memory. The same Linpack binary was used along with the same input parameters as described above. This yielded 131 total GFLOPS from the unit at a maximum power consumption of 521 Watts AC. **The 1U Twin™ therefore develops 251.5 GFLOPS / kW.** Stream Triad performance is 12 GBytes / Sec (6 GBytes / Sec for each board) at 351 Watts. Idle power consumption is 218 Watts AC. The power measurements were made with the Chroma unit referenced above and reflect the maximum power during the workload run. Table 1 compares the Linpack, Stream and Idle power consumption for the two units on a 'per server' basis.

	Linpack	**Stream**	**Idle**
SuperBlade®	221	155	98
1U Twin™	261	176	109

Table 1 : Comparison of Per Node AC Power Consumption

An interesting line of questioning that was explored further with the 1U Twin™ were the effects of low voltage, low power processors on the Linpack and Stream measurements. Figure 1 below shows an AC power trace for the 1U Twin™ populated with Xeon L5420 low voltage parts at 2.5 GHz (red line) and standard voltage 3 GHz parts (blue line). Note that the maximum power consumption reported for the low voltage parts running Linpack (521 Watts) is quite a bit higher than the average power, 509 Watts. The duration of the Linpack run used to calculate this average is represented by the yellow vertical lines on the chart. The total energy used to complete the calculation is only 1.5% greater for the standard voltage parts. This reflects the fact that the low voltage configuration operates at a lower power level, but takes longer to complete the calculation.

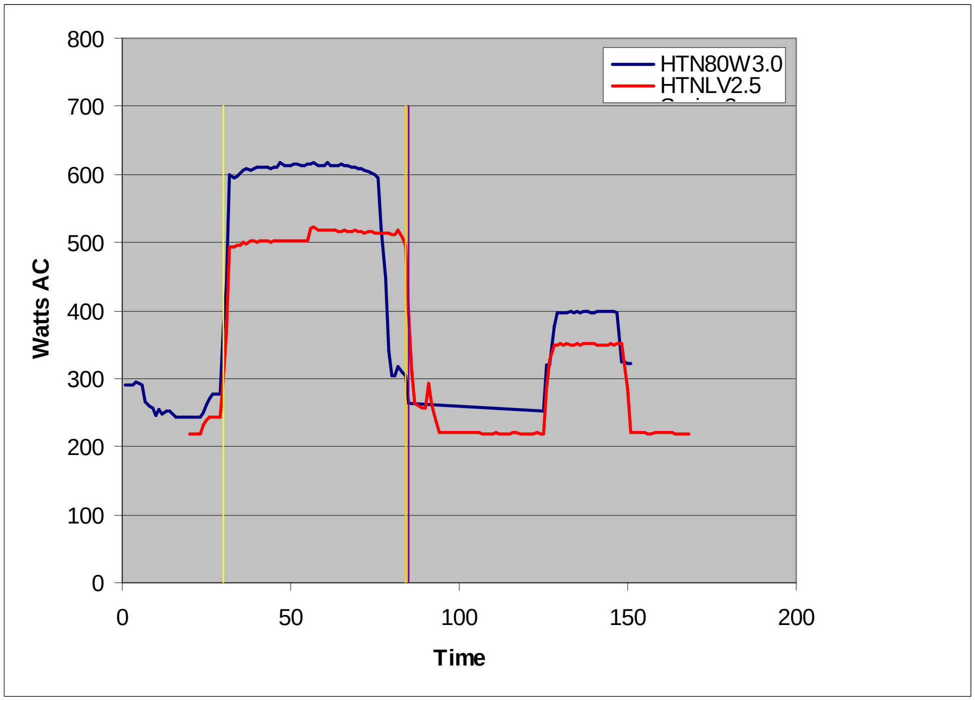

Figure 1: Power Consumption for 1UTwin™

The second bump in power in the traces, from 125 seconds to 150 seconds, is for the Stream workload. Here we see almost equal execution time for the two configurations as expected, given that Stream performance is largely independent of the CPU. Power consumption for the 3.0 GHz CPUs is significantly higher resulting in 10% greater overall energy used to complete the computation.

SPECjbb2005 and SuperBlade®

The SPECPower benchmark has not yet been specified for blade type systems. This prevents reporting SPECpower results for both SuperBlade® and 1U Twin™. In the interim we report the performance and power consumption for SuperBlade® while running the SPECjbb2005 workload. The SuperBlade® hardware was configured as already reported above with the following exceptions. Each blade was configured with 2x4GB DDR2-667 ECC DIMMs. The unit had two PSUs plugged in and delivering both power and air, while the third and fourth were mounted and delivering air via the integral fan units. The third and fourth units were not however delivering DC power. The fan speed control was set to 'Auto', which showed no appreciable difference from fan level = 1. On the software side, the Windows 2003 Server 64 bit operating system was installed. A freeware rsh implementation was installed to provide a scriptable mechanism for launching the SPECjbb2005 workload on each system from a single 'master' blade. The SPECjbb2005 v1.07 kit was installed on each blade along with the JRockit R27.5.0 JDK1.6.0_03 windows 64 bit JVM.

The workload was configured on each blade as follows: Four JVMs were run. The JVM command line options were JAVAOPTIONS=-Xms1600m –Xns1400m –Xmx1600m -XXaggressive -XXthroughputcompaction -XXcallprofiling -XXlargePages -XXlazyUnlocking -Xgc:genpar -XXgcthreads=2 -XXtlasize:min=4k,preferred=512k. Affinity was used to tie each JVM to a CPU as 03/0C/30/C0.

The workload was launched on each blade via rsh and batch files. The synchronization error across the blades was estimated at 5 seconds by timing the batch file on the master blade. The power usage for the whole unit is measured at 220V AC input with the Chroma unit noted above. Power was logged on both 5 second and 1 second intervals. The SPECjbb script reported valid runs on all 14 blades. Each run covered from 1 to 8 warehouses. The expected peak was at 2 warehouses and scoring was based on the 2, 3, and 4 warehouse results. A typical plot is shown in Figure 2.

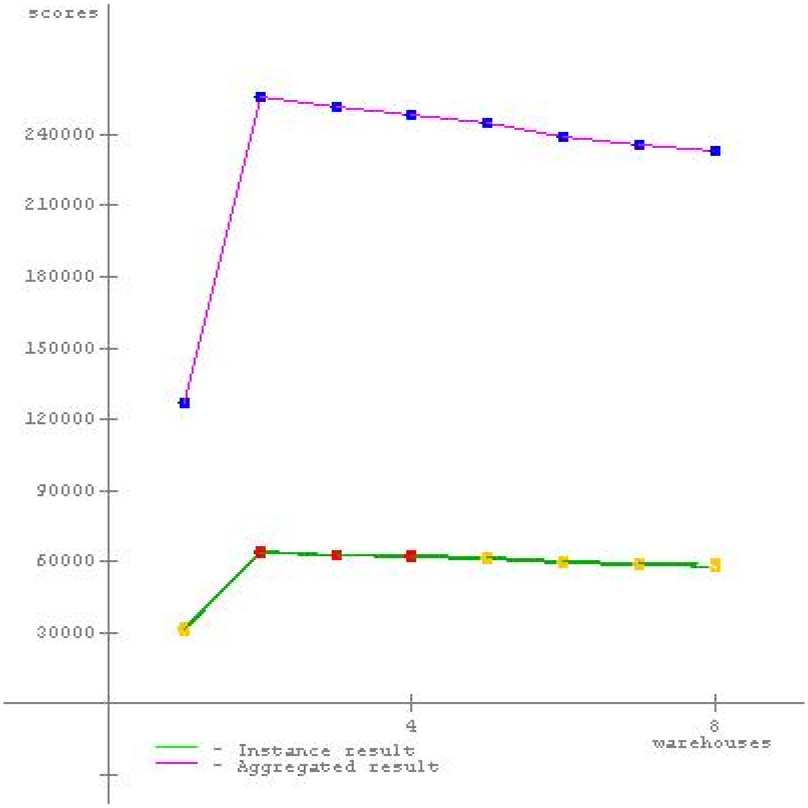

Figure 2: Typical SPECjbb2005 Run for a Single Blade

Each blade turned in an average score of 252.34K BOPs for a total of 3.533M BOPs for the 14 blade unit. The range of scores across blades varied by 2K BOPs (max score – min score).

A trace of the power usage is shown in Figure 3. Each data point represents the whole unit power usage on a 5 second reporting interval. The Chroma power meter was configured to report the average power over the 5 second interval. The vertical red lines are the estimated begin and end times for the scored interval of the workload run, e.g. warehouses = 2, 3, 4. The dips clearly delineate the warehouse number transitions.

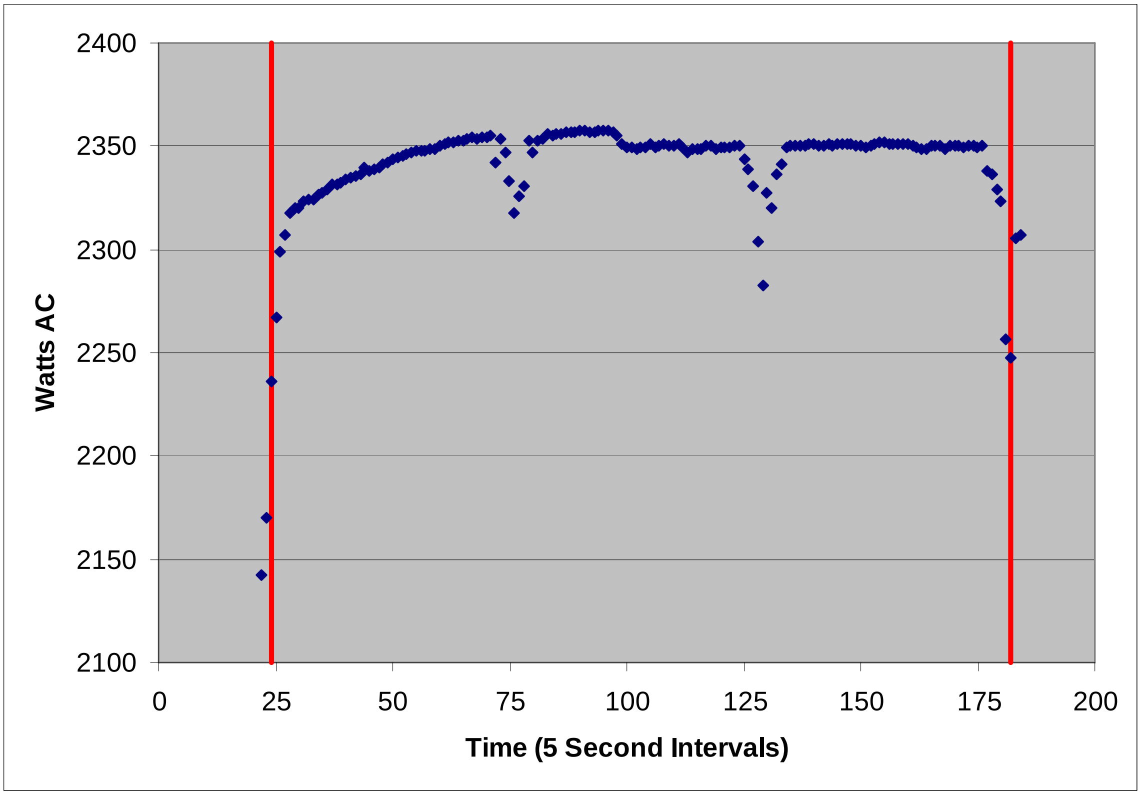

Figure 3: SuperBlade® Power Trace for SPECjbb2005 Workload

The start and end locations for the scored interval were placed based on the power dips and adjusted based on the known warehouse run time (240 seconds) and ramp up / down intervals (~23 seconds). The power usage for the scored interval was calculated in three metrics. The max metric (max power recorded between the vertical red lines) yielded 2358 Watts AC. The average metric (average across all data points between the vertical red lines) yielded 2342.7 Watts AC. This value is clearly too low as it includes the ramp up / down segments associated with the power dips which are not scored. The third metric represented an attempt to compute the average power usage only during the scored interval (240 seconds) for each warehouse. This came to just under 2346 Watts AC. A simpler alternative is to note that during the 4 warehouse run, the data points in Figure 4 all cluster around 2350 Watts. This value was chosen as the reported power usage for the scored run. The following values of interest can then be computed: **overall BOPs/Watt = 1503**, average power per blade during the scored run = 168 Watts. The Idle power was 1183 Watts total or 85 Watts per blade. The difference in this figure for Idle power and that given in Table 1 can be attributed to differences in the number of PSU modules delivering power which impacts the overall conversion efficiency. This effect is more pronounced at low loadings such as the Idle state.

Conclusion

A range of performance and power consumption tests were run on Supermicro's latest DP server products incorporating the Intel 5100 chipset. These products included the award winning SuperBlade® and 1U Twin™ servers targeted at data centers and HPC applications. Several workloads were used to exercise the systems including SPECjbb2005 to represent Internet-centric business applications and Linpack and Stream to represent HPC applications.

The results clearly indicate world class performance. The 1U Twin™ running Linpack yielded 131 total GFLOPS at a maximum power consumption of 521 Watts AC or **251.5 GFLOPS / kW.** SuperBlade® produced an average 65.65 GFLOPS per blade Linpack performance, for 919.1 total GFLOPS across all 14 blades. With a maximum recorded power consumption of 3100 Watts AC, this corresponds to **296 GFLOPS / kW for the SuperBlade®.**

Other performance measures were equally impressive. A highly optimized Stream binary run on the SuperBlade® yielded 6.0 GBytes / Sec on the Triad portion for a total of 84 GBytes / Sec across all 14 blades. Power consumption was measured at a max of 2175 Watts AC, or 155 Watts per blade. Stream Triad performance for the 1U Twin™ was 12 GBytes / Sec (6 GBytes / Sec for each board) at 351 Watts. Finally the performance and power consumption for SuperBlade® while running the SPECjbb2005 workload were measured as follows: **overall BOPs/Watt = 1503**, average power per blade during the scored run = 168 Watts.

The results from these performance measurements indicate that products such as the Supermicro SuperBlade® and 1U Twin™ are well suited to satisfy the increasingly stringent demands placed on modern data centers by both performance and energy efficiency requirements. These challenges must continue to be met for increased economic and social growth and development.

References:

[1] The $0.075 / kW-Hr figure is an average of the 2007 industrial electric power costs in 18 industrialized countries weighted by GDP. Data was obtained from the US Energy Information Administration, Monthly Energy Review, May 2007; and International Energy Agency, Energy Prices & Taxes- Quarterly Statistics; and International Monetary Fund, 2007.

[2] LBNL. 2006. High-Performance Buildings for High-Tech Industries, Data Centers. Lawrence Berkeley National Laboratory. http://hightech.lbl.gov/datacenters.html (accessed April 6, 2007).

[3] Greenberg, Steve, Evan Mills, Bill Tschudi, Peter Rumsey, and Bruce Myatt. 2006. "Best Practices for Data Centers: Lessons Learned from Benchmarking 22 Data Centers." Proceedings of the ACEEE Summer Study on Energy Efficiency in Buildings in Asilomar, CA. ACEEE, August. Vol 3, pp 76-87. http://eetd.lbl.gov/emills/PUBS/PDF/ACEEE-datacenters.pdf.

[4] "A Simple Model for Determining True Total Cost of Ownership for Data Centers", J. Koomey et. al., Uptime Institute. 2007. Site infrastructure capital costs can vary significantly according to desired Tier of functionality assumption.

[5] It is important to recognize that both the cost of electricity and the overhead factor for power distribution and cooling can vary by large factors. Thus the derived value should be thought of as a rough approximation to a median value of a distribution with a very large deviation. For example, the cost of electricity can vary from $0.025/kWHr to $0.20/kWHr depending on the geography supplier agreements. The overhead factor can very from a factor of 1.5 to perhaps 3 depending on the datacenter tier classification and details of the cooling scheme and geographic factors. The interested reader is advised to survey the references above and adjust the model to meet their unique situation.

[6] The cost of power distribution and cooling equipment as a fraction of the total cost of a new data center can vary widely depending on a number of factors. These include the data center tier classification, density, amount of attached office space and financial factors. The interested reader is advised to survey the Uptime Institutes whitepapers on this topic.