Datacenter Power Delivery Architectures : Efficiency and Annual Operating Costs

Paul Yeaman, V•I Chip Inc.

Presented at the Darnell Digital Power Forum
September 2007

Abstract

An increasing focus on datacenter efficiency, combined with the evolving trend of microprocessor core voltages below 1V requires an analysis of existing and proposed datacenter power delivery architectures. Additionally, multi-core architectures are increasing the number of cores per blade, thereby increasing overall power. Here, four distinct power delivery architectures will be analyzed and benchmarked in terms of overall efficiency, power conversion footprint, and annual electrical operating costs.

Introduction

Recent years have seen a continued surge in the number of active Datacenters containing potentially hundreds of thousands of low voltage, high current microprocessors. The overall power consumption of a datacenter can easily exceed 1MW. Power used by servers (incl. cooling and auxiliary infrastructure) was 1.2% of total US electricity demand in 2005 and cost $2.7 B[i] which can only increase as data volume doubles every 12-18months[ii]. Additionally, consolidation techniques such as virtualization - reaching 25% of x86-based servers by 2010[iii] - contribute to the spread of larger datacenters.

This high power consumption, especially in light of initiatives to increase overall datacenter efficiency, necessitates a study of efficiency within the rack system itself from the AC input to the rack to the low voltage high current loads (typically Microprocessors and Memory).

The ability to limit power consumption is further complicated by continuing evolution in the computing industry. Shrinking processes used in the fabrication are leading to lower microprocessor core voltages, with many product roadmaps indicating that by 2010 the core voltage will be at 0.8V[iv]. Also multi-core architectures are driving the total power consumption of a blade or motherboard higher, with new blades exceeding 1kW of power consumption within the next decade[v].

Problem Statement

From these two developments it becomes clear that a datacenter power delivery architecture must be both optimized for efficiency and to maintain high power density. For most power delivery architectures these two requirements are fundamentally opposed to each other: increasing efficiency increases the size of a power component and decreasing the size lowers the efficiency.

What follows is an analysis of four architectures for power delivery from the AC input to the rack to the sub-volt loads with benchmarks in terms of efficiency, power density, total cost of

ownership, and scalability. The benchmarks can be further described as follows:

1) *Efficiency* – Power delivered to a 1.xV 100+ Ampere load divided by input power at the AC plug.
2) *Power Density* – Total footprint of the power components in the datacenter system, including power supplies, DC-DC converters, and POL regulators.
3) *Total Cost of Ownership* – The total cost of operating the power delivery architecture based on the cost of the power dissipated in the power train (efficiency losses + distribution losses), the cost of removal of the heat from the environment (cooling system energy requirements).
4) *Scalability* – The ability of a system to be scaled up or down in terms of active power trains or active blades to match the active power delivery to the active processor elements.

For each of the following architectures, the system assumptions are summarized in Table 1.

Blade	
Loads	6 processor, 1.0V @120A ea.
	6 memory, 1.5V @50A ea.
	Misc. rails 12V @150W
Total Loads	1320W (~1032A on board)
Board Impedance	1.5mΩ (Regulator input current impedance)
Rack	
Number of blades	30
Distribution Impedance (AC-DC to blade)	2.0mΩ
Datacenter	
Duty Cycle /Rack	65%
Electricity cost	$0.14 kW/hr

Table 1: System Assumptions for architecture comparison

Some additional system assumptions are made which are mentioned below:

1) *Light load / no load operation is equivalent for each architecture*. For the purposes of the cost analysis, it is assumed that for each of the architectures discussed below, operation at light / no load results in equivalent loss, therefore no cost savings. In reality, power architectures which are scaleable will most definitely have better performance at light load, and therefore further cost savings.
2) *Low Voltage High current losses are negligible.* In reality, this is far from true, as the loss between regulator output and microprocessor or memory can be many watts/100A. However, in this analysis, there is no difference between architectures in this matter, and therefore no loss assumed.
3) *The blade does not include high current loads above 1.*5V. In actual practice, there may be loads at 1.8V, 2.0V, 2.5V, or 3.3V and for each of these outputs, the various architectures offer different balances in terms of cost savings.
4) *Redundant feeds / architectures are not considered.* This considers specifically a single power train originating at the AC input and terminating at a low voltage, high current load. There are no OR-ing losses, redundant feeds, or additional power trains included in this analysis.

Architecture Comparison

AC – 12 – 1.xV: Baseline Architecture

In this baseline architecture, AC power is converted to 12V DC using a bulk PFC supply in the rack and distributed throughout the rack to the various blades. On the blades the 12V power is regulated to the low voltage loads using conventional multi-phase buck regulators for the microprocessors and memory. The miscellaneous rail voltages are provided directly from the 12V blade input. Figure 1 illustrates the basic architecture.

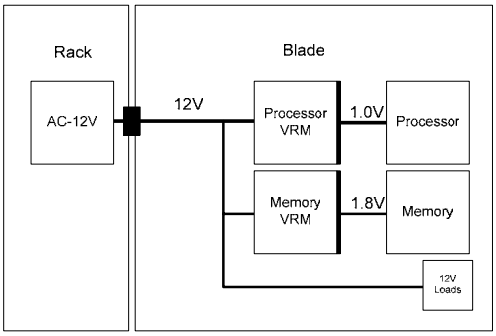

Figure 1: AC – 12V – 1.xV architecture

Figure 2 illustrates two approaches for implementing this in an actual system. In the first approach, a single bulk supply provides the 3819A for the 30 blades in the rack. This requires some method of bussing the 3kA+ around the system. The second approach breaks the single supply into multiple smaller supplies, each feeding a blade or cluster of blades. In this case, each supply needs only to supply a few hundred amps, depending on how it is partitioned. However, with the supplies located close to the blades, the blade space is now quite limited.

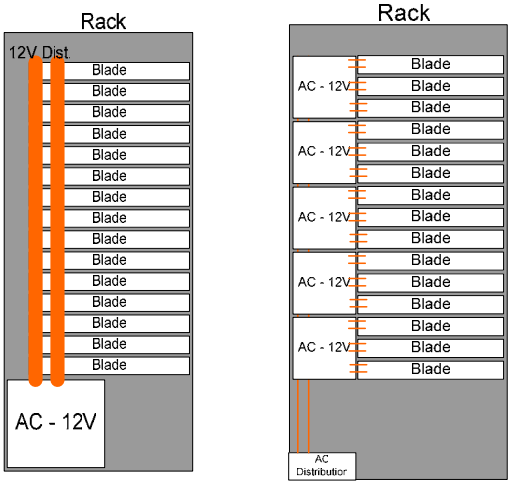

Figure 2: AC – 12 – 1.xV implementation options (not all 30 blades shown)

Based on the assumptions listed in Table 1 this architecture has the following attributes:

- Total losses due to distributing power from the AC power supply to the regulator inputs: 1.5kW (51.6W per blade)
- Total cost per year of operating the rack with the current power architecture (total cost of all power conversion + distribution losses + air-conditioning energy) : $19,770 ($659/blade)

AC – 384 – 12 – 1.xV

In the second architecture, shown in Figure 3, AC power is converted to a power factor corrected (PFC) 380V and distributed throughout the rack to the various blades. On the blades the 380V is converted to 12V using a 384-12V Bus Converter Module (BCM) [vi] and then regulated to 1.xV for microprocessors and memory. This architecture reduces the rack distribution losses from 32W per blade to less than a watt for the same distribution impedance. While it does add approximately 54W of dissipation to the blade itself, it decreases the power dissipation of the AC-DC power supply and substantially decreases the size, leading to overall gains in both size and efficiency. This is reflected in Figure 4, where the localized AC-DC power supply decreases in size by 40%, while the necessary blade area grows proportionally less, leading to a net gain in usable space on the blade itself.

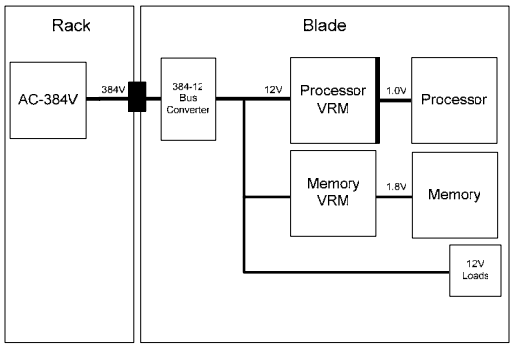

Figure 3: AC – 384 – 12 – 1.xV architecture

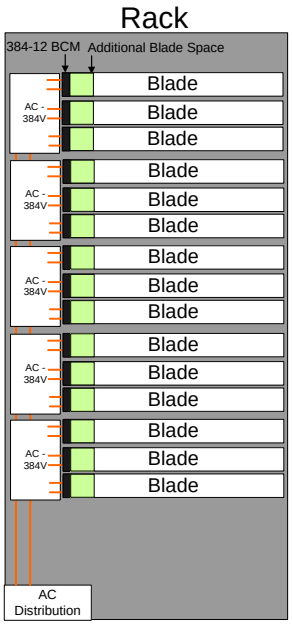

Figure 4: AC – 384 – 12 – 1.xV implementation

Based on the assumptions in Table 1, this architecture results in a total savings of $107 per blade and $3,208 per rack per year operation compared to the baseline architecture. Some additional benefits of this architecture:

- The space savings referenced earlier enable a net increase in usable space on the blade.
- The overall efficiency of the AC-12V power conversion increases by approximately 0.5% (reduction in losses)
- The introduction of a 384-12 300W conversion stage to a 1kW+ blade introduces a scaleable and controllable boundary to the architecture. Appropriate control of this boundary would enable optimization of light load efficiency.

AC – 384 – 48 – 1.xV

In this third architecture, AC power is converted to PFC 380V and distributed throughout the rack to the various blades. On the blades the 380V is converted to 48V using a similar bus converter to the previously discussed architecture and then directly converted and regulated to 1.xV for microprocessors and memory, using ZVS Buck – Boost Pre-Regulator Modules (PRMs)[vii] and Sine Amplitude Converter (SAC) Voltage Transformation Module (VTM)[viii]. In addition to increase power density and efficiency, the use of these power components in providing DC-DC conversion direct from 48V to low voltage enables the elimination of capacitance at the microprocessor (or memory) socket, resulting in a further increase in power density.[ix] This architecture is shown in Figure 5 and a potential implementation is shown in Figure 6.

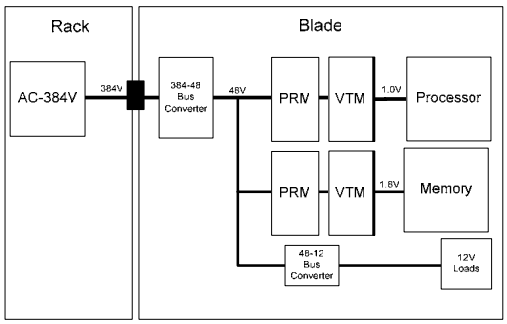

Figure 5: AC – 384 – 48-1.xV architecture

Increasing the blade distribution voltage to 48V now requires a 48-12 bus converter for the 12V auxiliary loads, resulting in lower efficiency for that particular load. However, this is more than compensated for by the fact that the board distribution losses drop from 19.2W to 1.1W with the increase in distribution voltage.

Additionally the efficiency of the 48-1.xV stage is approximately 5 percentage points higher than the 12 – 1.xV stage. Overall this results in a total savings of $260 per blade ($7,796 per rack) per year operation compared to the baseline architecture.

Just as with the previous 384V rack distribution architecture discussed, the capability exists for additional controlled scalability by selectively enabling and disabling the 384-48 converter based upon the load requirements or blade usage. This enables lower system loss at lighter loads due to reduced standby losses with idle converters.

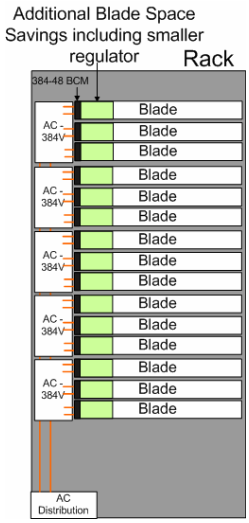

Figure 6: AC – 384 – 48 – 1.xV implementation

AC – 48 – 1.xV

In this fourth architecture, AC power is converted to 48V DC using a bulk supply in the rack and distributed throughout the rack to the various blades. This bulk supply can be a single supply or multiple smaller, localized supplies. On the blades the 48V is directly converted and regulated to 1.xV for microprocessors and memory. As in the previous example, a 48-12 bus converter is used for the auxiliary 12V rails. The system block diagram is shown in Figure 7, and two possible physical implementations of the system are shown in Figure 8.

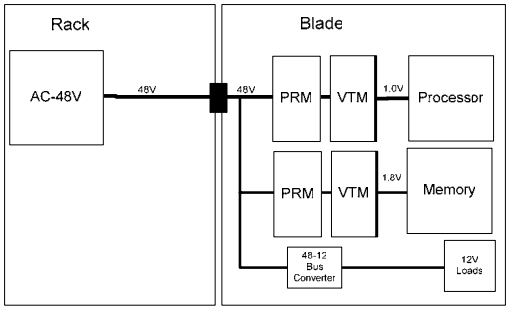

Figure 7 : AC – 48 – 1.xV architecture

While this architecture contains higher rack distribution losses compared to the 384V architectures, the projected losses given the impedance in Table 1 is 2W per blade (60W / rack) which is considerably smaller than the projected ~1kW/rack losses in the baseline architecture. Furthermore, this architecture takes advantage of the ~5 percentage point increase in utilizing 48V direct – to – load conversion compared to the 12V architecture.

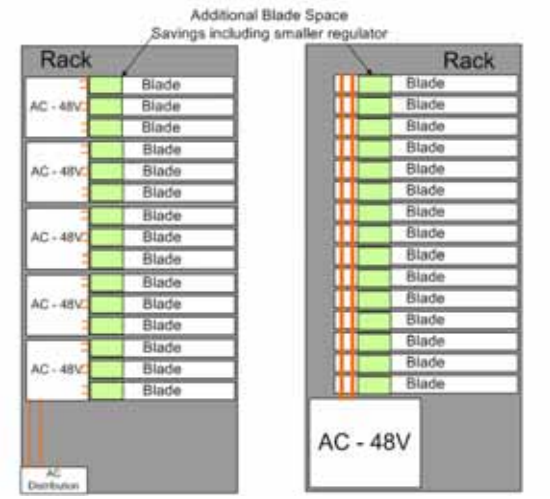

Figure 8: Possible implementations of an AC – 48V – 1.xV architecture

Overall, this represents a savings of $239/blade ($7,710 per rack) per year operation compared to the baseline architecture. In addition to the cost savings, this architecture provides the following attributes:

- AC – 48V AC-DC power supplies are standard products offered by a large number of different manufacturers, whereas the AC – 384 PFC supplies are considerably more limited in choices
- 48V distribution in a rack is a well known and characterized practice in central office telecommunication applications, whereas 384V distribution would be a new practice.
- 48V is a SELV voltage whereas 384V is a hazardous voltage. Distribution of a SELV voltage greatly simplifies the distribution, shielding, and electrical connection of the blades to the rack.

General Architectural Comments

AC – 384 – 12 – 1.xV and AC – 384 – 48 – 1.xV architectures utilize on-blade high voltage to low voltage conversion. This essentially changes the AC-DC supply from a single multi-kW bulk power supply to a multi-kW PFC supply followed by an array of blade level 300W 384-12 or 384 – 48V converters. Thus half of the AC-

DC supply function is moved onto the blade itself.

This partitioning enables a new layer of digital power management in which bulk power is controlled at the blade level. The multi kW bulk supply is now broken into smaller blade level conversion stages which can be disabled as required by blade power requirements and can be controlled independently of the AC-DC conversion stage.

The partitioning of blade functionality into memory and processor clusters essentially results in ~2:1 ratio of processor/memory to converter. Thus if a pair of processors on a blade is to be disabled, or run in power save mode, the digital control of that pair can be tied directly to a 384-12 or 384-48 converter. This will have the greatest impact if the blade is less than 30% loaded.

In the two architectures utilizing distributed 384V throughout the rack, the opportunity for further efficiency gains exists in Datacenter architectures with a distributed 384V system[x]. This would enable the AC – 384V stage to be bypassed and the blades fed directly from the datacenter distribution busses. The partitioning and control capability mentioned above would have an even greater impact on light load operation.

Conclusion

Four architectures for converting and regulating power from the AC input of a 30 blade datacenter rack have been analyzed in terms of total cost of ownership, efficiency, and density. In terms of highest overall system power density, efficiency, and total cost of ownership, the AC – 384 – 48 – 1.xV solution provides the maximum efficiency, annual cost savings, and smallest power solution footprint. A comparison of the cost savings of the four architectures is shown in Figure 9.

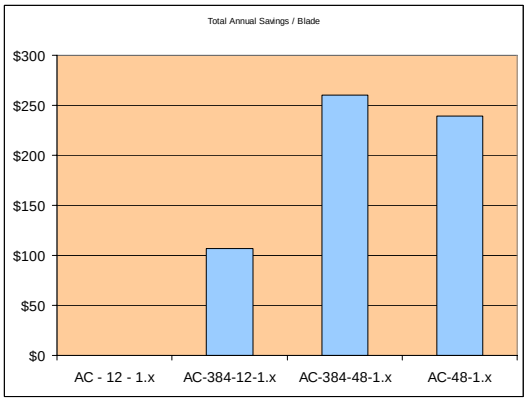

Figure 9: Annual Blade savings comparison

The datacenter power delivery architecture has a direct impact on the cost of operating the datacenter itself, as well as the options enabled by the use of power components in addition to, or in place of power supplies. The choice of architecture should represent the constraints imposed by the installation in conjunction with the optimal efficiency, performance, and size available.

---

[i] Estimating total power consumption by servers in the US and the world; Koomey, Lawrence Berkeley National Laboratory, February 15, 2007

[ii] Building Your Virtual and Blade-Based Infrastructure, Bob Kohut, HP, Jake Smith, Intel & Stephen Shultz, VMware, Inc. searchservervirtualization.com

[iii] IDC Server virtualization Projections Sep '06 and Feb '07 update.

[iv] PSMA Power Technology Roadmap 2006, pp. 32. Power Technology Roadmap for Microprocessor Voltage Regulators, Ed Stanford, Intel.

[v] PSMA Power Technology Roadmap 2006, pp. 23-24. Power Trends – High Performance Servers. Shaun Harris, H-P.

[vi] V•I Chip B384F120T30 Bus Converter Module datasheet

[vii] V•I Chip P045F048T32AL datasheet

[viii] V•I Chip V048F015T100 datasheet

[ix] "High Current Low Voltage Solution For Microprocessor Applications from 48V Input." Paul Yeaman, *PCIM 2007 proceedings.*

[x] "DC Power for Improved Datacenter Efficiency", Ton (Ecos), Fortenbery (EPRI) & Tschudi (Lawrence Berkeley National Labs), January 2007.