データセンターにおける電力伝送アーキテクチャ：効率と年間運用コストの比較について

Paul Yeaman, V·I Chip Inc.

要旨

マイクロプロセッサのコア電圧が1V以下へ進化を遂げる中、データセンターの効率化に注目が集まっている。データセンターの効率化を図るには、既存および現在提案されているデータセンター電力伝送アーキテクチャの分析が必要である。さらに、マルチコアアーキテクチャによって1ブレードあたりのコア数は増加傾向にあり、それにより全体の電力消費量はさらに増加することになる。ここでは、全体的な効率、電力変換部分の設置面積、および年間の電気的な運用コストの点から、4つの異なる電力伝送アーキテクチャを分析し、ベンチマークを行う。

はじめに

近年、何十万もの低電圧大電流仕様マイクロプロセッサを備えたデータセンターの数が増え続けており、1つのデータセンターが消費する総電力量は1MWを軽く超える。2005年にサーバーが消費した電力（冷却設備と補助装置を含む）は米国の総電力需要の1.2%を占め、電力コストは27億ドルに達したi。また、データセンターで処理されるデータ量は12～18ヶ月に2倍のペースで増えているため、それに伴って電力コストも増加することが予想されるii。さらに、2010年までにx86ベースのサーバーの25%に達するiiiといわれる仮想化のような統合化技術の普及が、データセンターの大規模化の一因になっている。

この大電力消費を改善するには、データセンターの全体的な効率向上の第一歩として、ラックへのAC入力から低電圧大電流負荷（主にマイクロプロセッサとメモリ）までのラックシステム自体の効率を調査する必要がある。

コンピュータ業界における演算能力の進化によって、電力消費量を制限することはさらに困難になる。数多くの製品ロードマップでは、ファブリケーション工程で用いられるシュリンク技術によってマイクロプロセッサのコア電圧がさらに低減化され、2010年までにコア電圧が0.8Vivになることが示されている。また、マルチコアアーキテクチャは、これから10年以内に消費電力が1kWを超える新型ブレードが登場するように、ブレードやマザーボードの総電力消費量をさらに高く押し上げているv。

問題提起

このようなコア電圧の低電圧化とマルチコア化という2つの技術進歩から考えて、データセンターの電力伝送アーキテクチャでは、効率を最適化すると同時に高電力密度を維持しなければならないのは明らかである。しかしながら、効率を高めようとすると電力コンポーネントのサイズは大きくなり、電力コンポーネントのサイズを小さくすると効率は低下するため、ほとんどの電力伝送アーキテクチャにおいて、効率の最適化と高電力密度の維持という2つの要件は基本的に相反するものである。

ここでは、ラックへのAC入力から低電圧負荷までの電力伝送について、効率、電力密度、総所有コスト、および拡張性の観点から、ベンチマークを使って以下の4つのアーキテクチャを分析する。ベンチマークについては、以下のように説明される。

1) 効率 - 1.xV 100A～の負荷に供給される電力を、ACプラグでの入力電力で除算する。
2) 電力密度 - 電源装置、DC-DCコンバータ、およびPOLレギュレータといったデータセンターシステム内の電力コンポーネントの設置面積の総計。
3) 総所有コスト ― 電力を伝導する経路（パワートレイン）で消費する電力のコスト（効率損失＋配電損失）および周囲からの熱を除去するコスト（冷却装置の電力需要）に基づいて、電力伝送アーキテクチャを運用するための総費用を算出する。

4) 拡張性 – 現在使用しているプロセッサ素子に電力伝送アーキテクチャを適合させるために、パワートレインやブレードについて拡張／縮小するシステムの能力を示す。

次の各アーキテクチャについて、システムの前提条件を表1に要約する。

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: アーキテクチャ比較のためのシステムの前提条件

さらに、以下に示すような別の前提条件を立てる。

1) 軽負荷／無負荷での動作は各アーキテクチャで同等とする。後述の各アーキテクチャについて、コスト分析の目的で軽負荷／無負荷で動作させると、結果的に等価損失になるため、コスト削減にはならない。実際には、拡張性のある電力伝送アーキテクチャは、軽負荷においてより良い性能を示し、さらにコスト削減効果が得られると思われる。

2) 低電圧大電流での損失はわずかとみなす。実際には、レギュレータ出力とマイクロプロセッサ／メモリの間の損失は100Aでは大きな電力になりえるので、これは正しいとはいえない。しかしながら、この分析の場合、本前提条件において各アーキテクチャ間で違いがないため、損失はほとんどないとみなされる。

3) ブレードは1.5V を超える電圧での大電流負荷を持たない。実際の運用では、1.8V、2.0V、2.5V、もしくは3.3V での負荷があるかもしれず、そして、これらの出力のそれぞれについて、コスト削減の観点から各種アーキテクチャによってさまざまな出力機能が提供されている。

4) 冗長給電／アーキテクチャを考慮しない。この分析では、AC入力で動作を開始し、低電圧大電流負荷で動作を終了するシングルパワートレインとみなす。また、ORingによる電力損失、冗長給電、またはパワートレインの追加は含まないものとする。

アーキテクチャの比較

AC - 12 - 1.xV: Baseline Architecture
（本構成がベンチマークの基準となる）

この基準アーキテクチャでは、ラック内のPFC（力率改善回路）を使った大容量電源によってAC電力が12V DCへ変換されてから、ラック全体の各ブレードへ配電される。ブレードでは、マイクロプロセッサおよびメモリ用の一般的なマルチフェーズ降圧レギュレータを使って12Vの電力が低電圧負荷へ定電圧化され、小電力負荷用レール電圧は12Vブレード入力から直接供給される。図1に基本アーキテクチャを示す。

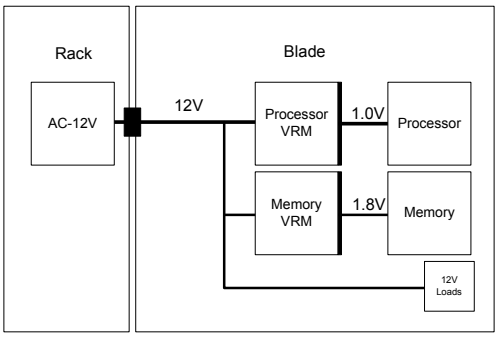

Figure 1: AC – 12V – 1.xV architecture

このアーキテクチャをシステムに実装するための2つのアプローチを図2に示す。最初のアプローチでは、単一の大容量電源からラック内の30個のブレードへ3819Aが供給される。この場合、システム全体に渡って3000Aを超える電流を伝送する手段が必要になる。2番目のアプローチでは、単一の大容量電源が複数の小規模な電源へ分割され、それらの各電源から単一ブレードまたはブレードクラスタへ給電される。この場合、分割方法にもよるが、各電源は数百アンペアの電流しか必要としない。ただし、各電源はブレードの近くに設置されるため、ブレードの設置場所が非常に制限されることになる。

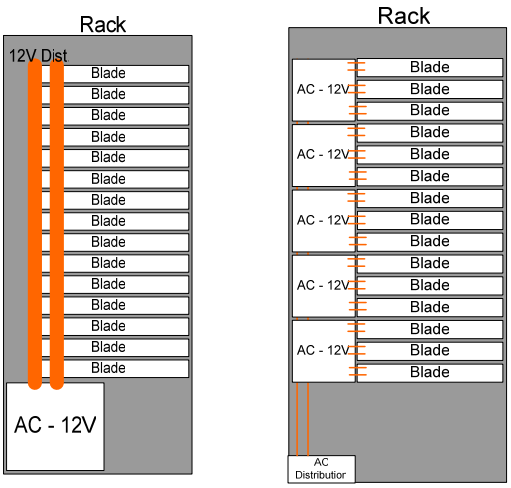

Figure 2: AC – 12 – 1.xV implementation options (not all 30 blades shown)

表1に示す前提条件に基づき、このアーキテクチャは以下の特性を持つ。

・AC電源装置からレギュレータ入力への配電によって生じる総損失量：1.5kW（1ブレードにつき51.6W）
・現在の電力アーキテクチャでラックを運用する場合の年間総費用（全電力変換の総費用＋配電損失＋空調エネルギー）：19,770ドル（1ブレードにつき659ドル）

AC – 384 – 12 – 1.xV

図3に示す2番目のアーキテクチャでは、AC電力がPFC（力率改善回路）によって380Vへ変換され、ラック全体の各ブレードへ配電される。ブレードでは、384-12V BCM（バスコンバータモジュール）を使って380Vが12Vへ変換されてから、マイクロプロセッサおよびメモリ用に1.xVへ定電圧化される。このアーキテクチャの場合、同じ配電インピーダンスで、ラックでの配電損失は1ブレードにつき32Wから1W以下へ減少する。また、ブレード自体に約54Wの損失が増える一方で、AC-DC電源装置の消費電力は減少し、サイズが大幅に削減されるので、サイズと効率の両面で全体的な効果が得られる。図4に示すように、分散型AC-DC電源装置のサイズが40%削減ことに比例してブレードの占有面積の割合は減少するため、ブレードの利用可能スペースが実質的に拡大されることになる。

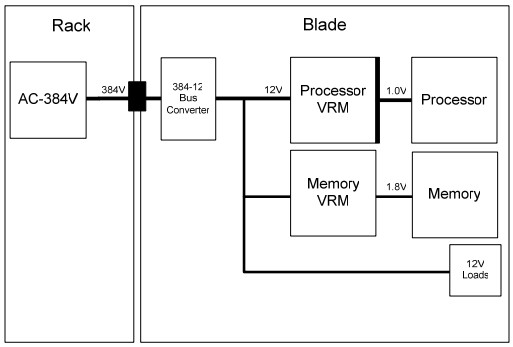

Figure 3: AC – 384 – 12 – 1.xV architecture

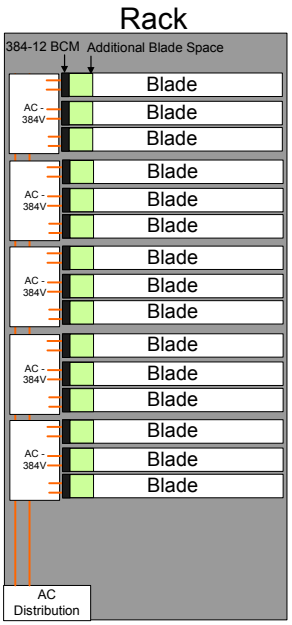

Figure 4: AC – 384 – 12 – 1.xV implementation

表1に示す前提条件に基づき、このアーキテクチャを基本アーキテクチャと比較すると、1ブレードにつき年間107ドル（1ラックにつき年間3,208ドル）の運用コスト削減になる。さらに、このアーキテクチャには以下のようなメリットもある。

・前述のような省スペース化により、ブレードの利用可能スペースが実質的に拡大する。
・AC-12V電力変換全体の効率が約0.5%高くなる（損失の減少）。
・300Wの384-12V変換を1kW超のブレードに採用することにより、拡張性および境界の制御可能性をアーキテクチャへ取り入れることが可能になる。この境界を適切に制御することにより、軽負荷時の効率を最適化できる。

AC – 384 – 48 – 1.xV

この3番目のアーキテクチャでは、AC電力がPFC（力率改善回路）によって380Vへ変換されて、ラック全体の各ブレードへ配電される。ブレードでは、前述のアーキテクチャと同様にバスコンバータを使って380Vが48Vへ変換されてから、ZVS Buck – Boost PreRegulator Modules (PRM)viiおよびSine Amplitude Converter (SAC) Voltage Transformation Module (VTM)viiiを使ってマイクロプロセッサとメモリ用に1.xVへ直接変換、定電圧化される。48Vから低電圧へ直接DC-DC変換するための電力コンポーネントを使うことにより、電力密度と効率の向上に加えて、マイクロプロセッサ（またはメモリ）ソケットでのバイパスコンデンサを除去でき、結果として電力密度をさらに高めることが可能になるix。

このアーキテクチャを図5に示し、実装例を図6に示す。

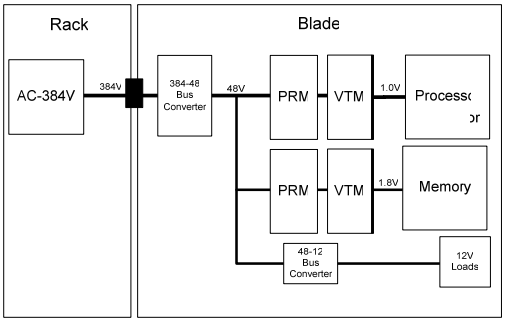

Figure 5: AC – 384 – 48-1.xV architecture

ブレードへの配電電圧を48Vへ増やすと、小電力負荷用12Vの48-12バスコンバータが必要になるため、その特定の負荷によって結果的に効率が低下する。ただし、配電電圧が高くなると配電損失は19.2Wから1.1Wへ低下するため、それにより効率の低下が補われることになる。
さらに、48-1.xV変換ステージでの効率は12–1.xV変換ステージに比べて約5%高くなるため、結果として、基本アーキテクチャと比較して1ブレードにつき年間260ドル（1ラックにつき年間7,796ドル）の運用コスト削減になる。
拡張性の制御に関しては、前述の384Vラック配電アーキテクチャと同様に、負荷要件やブレードの使用法に応じて384-48Vコンバータをイネーブル／ディセーブルの状態に切り替えられる機能が提供されている。

この機能を使ってスタンバイ状態のコンバータで生じるスタンバイ損失を低減することにより、さらに軽い負荷でシステム損失を低く抑えることが可能になる。

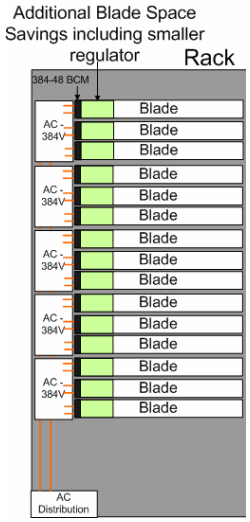

Figure 6: AC – 384 – 48 – 1.xV implementation

AC – 48 – 1.xV

この4番目のアーキテクチャでは、ラック内の大容量電源を使ってAC電力が48V DCへ変換され、ラック全体の各ブレードへ配電される。この大容量電源は、単一の電源としても、複数の小規模分散型電源としてもかまわない。
ブレードでは、48Vがマイクロプロセッサとメモリ用に1.xVへ直接変換、定電圧化される。前述の例で示すように、48-12バスコンバータが12Vの補助電圧用に使用される。システムのブロック図を図7に示し、2つの物理的な実装例を図8に示す。

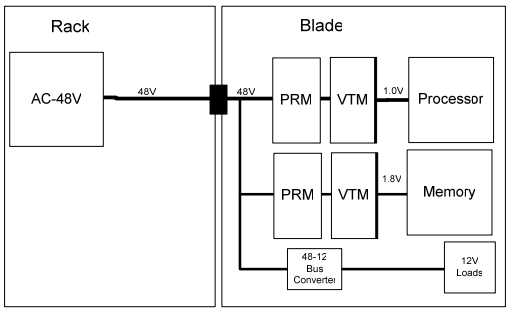

Figure 7 : AC – 48 – 1.xV architecture

このアーキテクチャでは、ラック配電損失が384Vアーキテクチャに比べて高い一方で、表1に基づくインピーダンスから算出される損失は1ブレードにつき2W（1ラックにつき60W）で、1ラックにつき最大1kWという基本アーキテクチャでの予想損失に比べて大幅に減少する。さらに、このアーキテクチャでは48Vの電力が変換されて直接負荷へ供給されることにより、12Vアーキテクチャに比べて効率が最大5%向上する。

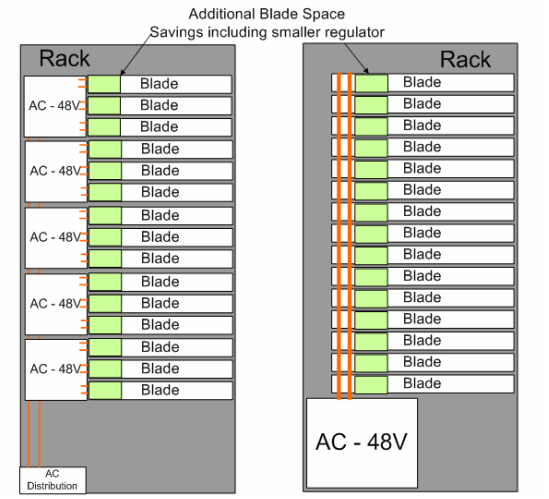

Figure 8: Possible implementations of an AC – 48V – 1.xV architecture

これは、基準アーキテクチャに比べて、全体として1ブレードにつき年間239ドル（1ラックにつき年間7,710ドル）の運用コスト削減になることを意味する。コスト削減に加えて、このアーキテクチャは以下のような特性を有する。

・AC - 48V 出力のAC-DC電源は数多くのメーカーから販売されている標準品だが、AC - 384 PFC電源の場合は選択の幅が非常に限定される。また、ラック内での48V配電は電話局の通信アプリケーションで既に実績のあるよく知られた方法だが、384V配電はまだ実績のない新しい方法であると思われる。

・48VはSELV電圧だが、384Vは危険電圧である。SELV電圧の配電では、配電、遮へい、およびラックへのブレードの電気的接続が大幅に簡略化される。

アーキテクチャ概評

AC - 384 - 12 - 1.xVおよびAC - 384 - 48 - 1.xVアーキテクチャでは、ブレード上で高電圧から低電圧への変換を行う。これにより、AC-DC電源は単一のマルチkW大容量電源からマルチkWのPFC電源へ実質的に変わり、それに続いてブレードレベルで300Wの384-12V変換または384 - 48V変換が行われる。つまり、AC-DC電源が果たす役割の半分はブレードへ移ることになる。

この機能配分により、ブレードレベルで大容量電力を制御するといった新しいデジタル電源管理のステージが可能になる。このように、マルチkW大容量電源の役割をより小さなブレードレベルの変換ステージへ分けることにより、ブレードの電力要求に応じてコンバータをディセーブル状態に切り替えることや、AC-DC変換ステージでの独立した電力制御が可能になる。

ブレードの機能をメモリやプロセッサクラスタへ配分すると、プロセッサ／メモリとコンバータの機能配分の比率は基本的に2対1になる。したがって、あるブレード上の一組のプロセッサがディセーブル状態か、あるいは省電力モードで動作している場合、その一組のプロセッサのデジタル制御を384-12または384-48コンバータに直接連動させることができる。これは、ブレードの負荷が30%より小さい場合に最も効果的と考えられる。

分散型の384VシステムXを利用するデータセンターでは、ラック全体に渡って384Vを配電するこれら2つのアーキテクチャを利用することによって、さらに効率を向上できる可能性がある。また、AC - 384V変換ステージをバイパスしたり、データセンターの配電バスからブレードへ直接給電したりすることも可能である。以上のような機能配分や制御能力は、軽負荷での運用時に一層高い効果が得られると考えられる。

結論

ここでは、総所有コスト、効率、および電力密度の観点から、30個のブレードを搭載したデータセンターのラックへ入力されるAC電力を変換および定電圧化する4つのアーキテクチャを分析した。
システム全体の電力密度、効率、および総所有コストの点から、AC -384 - 48 - 1.xVアーキテクチャが、最大の効率と年間コスト削減を実現し、電力コンポーネントの設置面積を最小化するソリューションを提供できることが分かる。4つのアーキテクチャのコスト削減効果を比較したものを図9に示す。

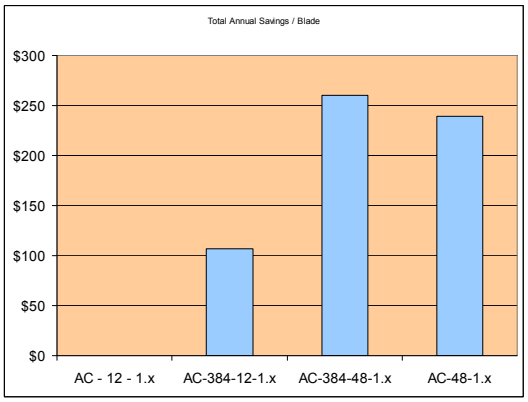

Figure 9: Annual Blade savings comparison

データセンター電力伝送アーキテクチャは、電源に加えて（または、電源の代わりに）使用する電力コンポーネントのオプション機能を利用できるという利点があると同時に、データセンター自体の運用コストに直接的な影響を与えるため、アーキテクチャを選択する際は、最適な効率、性能、および利用可能なサイズと併せて、設置によってどのような制約が課せられるかを考慮する必要がある。

---

[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.