

The Power System Challenge - Understanding the Total Picture

Randy Malik

Research Triangle Park, North Carolina

Email: Rmalik@us.ibm.com

I. Introduction

The power supply for a server system is not just a power supply but a system by itself. The most important requirement of a server system is its availability to the customer. Downtime of a few hours could cost a company millions of dollars. To keep a server system available for 100% of the time, full system redundancy is required. A complete server system consists of an AC source, AC/DC power supply, DC/DC regulators, system controllers, microprocessors, memory, I/O cards and storage devices. Fault in any one of these functional blocks could crash the whole system. It is, therefore very essential that redundancy is provided from AC input to all the way to the system boards. The power system is one of such critical functional blocks required to provide redundancy. Therefore, the power system architect has to understand the system in great detail to achieve 100% availability at a reasonable cost. For mission critical applications where a system handles real time commands and human lives are at stake, full system redundancy is mandatory. However, if the system application is for data processing, travel reservations, banking operations, wide area network communication etc., a fully redundant system could be an option because of high cost of redundancy. This paper details some of the trade offs such as cost, reliability, custom design and distributed power architecture to achieve a reasonable system availability.

II. Brief Description of Server Systems

Enterprise Server is a system designed to serve the need of large corporations, for example, banks, ATMs, insurance companies, reservations, travel and payroll etc. These are considered mission critical applications where an hour's downtime could cost millions of dollars in lost revenue. Almost all of these servers are fully redundant from AC to system boards. The host machine is generally an entire rack where multiple symmetric parallel processors process the information and store the transactions in a separate storage rack.

The power system consists of redundant AC sources, AC/DC regulators and DC/DC regulators. The AC/DC regulator output could be 350 V DC, -48V DC or some other convenient voltage. Almost all DC/DC regulators are redundant with isolation diodes or MOSFETs for fault protection. Systems with 350V DC distribution use a separate assembly for redundant DC/DC regulators while systems using - 48V power distribution use isolated point of load (POL) DC/DC regulators. AC power to the system rack is generally provided by two independent utility company feeds.

Mid Range Server is a system designed to serve the need of a midsize company application, for example, Internet service, Automated Assembly Line, e-mail service, and Local Area Network etc. Such applications require high system availability. For example, according to "Uptime Institute Standard", downtime of less than five minutes per year for the power system is considered very satisfactory. The AC could be available from either two independent AC sources or by one AC source and the other by an Uninterruptible Power Source. The AC/DC regulators are either redundant multiple output power supplies or single output AC/DC regulators followed by DC/DC regulators which may or may not be redundant.

Low End Server is a system designed for small business applications, for example an office, a warehouse or a store. The server could be as simple as a computer with a network card to enable communication with other computers. The power system for such an application is generally not redundant and, in most cases no UPS is required. A low end server has bare minimum functions and the data is generally backed up each day by either an external tape drive or an external bank of Hard Disk Drives. The system cost is the primary concern to the customer and in a free competitive market, the customer sets the price.

III. Power System Issues

1. **Technology Trend** for microprocessors, chip-sets, memory, I/O boards and Hard Disk Drives etc. to improve performance requires lower DC voltages and higher di/dt operation. The new memory architecture known as Double Data Rate (DDR) requires one of the DC/DC regulator to source and sink the current and tracking the memory bus voltage at the same time. The microprocessor voltage which is currently close to one volt needs very tight regulation and extremely small ripple. These requirements are forcing use of point of load regulators (POL) which have to be designed for very fast transient response.
2. **Power Density** in itself has no direct value for the customer. The main objective of a system designer is to allocate maximum space for system functions such as storage, computing, I/O cards etc., because the return on investment per cubic inch for system functions is much higher than the power system. Smaller the space required by the power system, more space available for the system functions.
3. **Power System Cost** could be a significant percentage of the overall system cost. This could vary from 5% all the way to 10% of the system cost. Therefore it is very crucial for the system designers to pay attention to the cost of a power system. A poor decision by the designer could result in either a costly power system or a system with high failure rate and worst of all a product recall.
4. **Time to Market** is the most significant factor in determining the power system architecture. While the time to qualify a custom designed power system still remains approximately nine months, while for servers, it has gone down from a high of 18 months to 24 months all the way down to six months! At a time when the power industry is still struggling with common power building block approach, the server industry has been using the functional block approach for a while to reduce time to market. The custom designed power system still remains the gating item.
5. **Power Engineering Resource** is becoming a scarce commodity as more and more system design companies continue to consolidate their resources for a core business. The server power system is not simply a power supply and requires a highly specialized team to optimize a reliable power system at a reasonable cost. An OEM organization generally lacks power system specialists who can really understand the application as well as the power system. Therefore, lack of a balanced system and power expertise could be very risky for an OEM to provide a satisfactory solution.

IV. Standard or Custom Power

There are various factors to be considered before deciding whether to use a standard or a custom designed power system. Some of the factors which could influence the decision one way or the other are described in the following paragraphs;

1. **Schedule** is one of the most important factors in determining success or failure of a product. If a product is late for the market because of schedule slip, it could have serious consequences for the bottom line of a company. It is a well known fact in the power industry that a custom designed

power system for a server application will need at least nine months to qualify. If the time to market for a product is less than nine months, the system designer is better off using existing power supplies or standard power modules.

2. **Cost** becomes an important factor if the server system is being sold in a competitive market. In such an environment, cost is the king. If a product does not meet a minimum cost target, it becomes very hard to sell it in the market place. The power system is generally 5% of the system cost, and therefore any saving in a power system could go a long way. However, if the cost of development compared to the savings achieved by custom design are insignificant and the life time demand is in thousands, custom design could be a better alternative.
3. **Reliability:** A custom designed product is generally less reliable than a standard product because of the learning curve. A standard product could very well be in the third or fourth generation of design/manufacture, and therefore all sorts of design and process problems are way in the past. Product recall and litigation issues should be carefully evaluated when considering a custom designed product.
4. **Flexibility:** In today's market environment, the product life span is generally 18 months to 24 months. Therefore, a custom designed product has very short, therefore the development and qualification costs are very high. In addition, there is hardly any time to improve the existing design but there is always pressure to reuse the custom power supply product in the next generation product.
5. **Design Roadmap:** Standard power products have different generation road maps for

design improvements. There are cost savings to be achieved from one generation to the other in such standard products. Such is not the case for a custom designed product.

V. Imbedded or Plug-in DC/DC Regulator

In a Distributed Power Architecture, point of load (POL) regulators are needed to provide power to individual functional blocks of a system. The point of load (POL) regulator could either be plugged into a socket or be a part of the system card. Alternatively, a POL regulator could be soldered on the system board. Each approach has its own merits and demerits depending on the application.

Advantage:

If there is enough board space, the imbedded solution could provide some cost advantage over a plug-in or a soldered in solution. This assumes only the cost of components and does not include the board space, manufacturing or test costs.

VI. Disadvantages:

1. **High Risk:** An embedded design is very high risk due to lack of design expertise and test time which can result in expensive rework, recall or worst of all catastrophic failure that could impact very expensive components such as microprocessors. The plug-in DC/DC regulators are designed by industry experts and have much lower risk of recall or rework.
2. **Infant Mortality:** A Plug-in DC/DC regulator is pre-tested and burned-in before being used in a system, on the other hand an imbedded regulator is tested with the system only. The imbedded solution without pre-testing and burn-in can have infant mortality problems resulting in rework of very complex boards.

3. **Total Product Cost :** The initial cost of the imbedded solution appears to be lower but the total expense over the life time of the product could be equal to or even higher than the plug-in solution depending on the product quantities shipped and the warranty costs associated. In addition, if there is a rework or recall, the cost could quickly run into millions of dollars.
4. **Flexibility:** A plug-in DC/DC regulator has a road map for improvement from generation to generation using better and more cost effective components without impacting the system board layout. For example, if the system is upgraded with a next generation microprocessor, there will be minimal impact on the board layout if a plug-in solution is used.
5. **System Board Issue:** The imbedded solution needs a fixed XY area on the system board and the ZY and ZX space does not get utilized. A plug-in module is designed with four to six 2 oz layers of copper, but the imbedded solution is allocated only two 1 oz copper layers. This causes loss of efficiency, and heat dispersal becomes a problem. In addition, the plug-in module stands up and has a better cooling in a forced air cooling environment. Also, in an imbedded solution, all of the system board layers are not available for high frequency paths, the path inductance is higher and therefore higher voltage spikes are developed across switching devices. Worst of all, the fast switching transients in an imbedded solution could interfere with the other system signals crisscrossing the system board.

VII. Power Systems : Redundant, Non-Redundant and Semi-Redundant

For mission critical applications, a system is required to be fully redundant. A fully redundant power system provides protection against single

point faults within the power system, but any short on the external power system bus could still crash the system. A system is considered fully redundant if the system architecture shall not allow the fault propagation from one system to the other.

The most common cause of downtime in a system is generally the failure of utility AC power. Some examples of AC power quality are given below;

Factories	= 0.9999	(four 9's)
Hospitals	= 0.99999	(five 9's)
Banks	= 0.999999	(six 9's)
Online Markets	= 0.9999999	(seven 9's)

For example, five 9's is equated to not more than 5 minutes of downtime per year.

VIII. Availability Issue

According to "The Uptime Institute (UTI)", a downtime of 5 minutes or less due to power system is considered satisfactory. This is equated to power availability of 0.99999 (five 9's). A fully redundant power system with redundant AC sources shall have no problem in meeting this requirement. However, many of the server systems can not afford the extra cost of such redundancy. For such systems, the most cost effective way to provide reasonable availability is to have redundancy for those functional blocks whose failure rate is very high, the AC source and AC/DC regulators are few of such examples where there is a high probability of failure because of the nature of power quality and the operating environment. It is for these reasons that redundant utility AC sources and the redundant AC/DC regulators are required for highly available systems. In mission critical applications such as enterprise servers, the redundancy is further extended to DC/DC regulators.

AC Redundancy: Two independent utility AC feeds applied to an N+N power system will provide the highest availability even if the individual source is of poor power quality.

For example, an AC source with 0.9999 (four 9's) availability (downtime 50 minutes/year) in conjunction with another similar independent AC source will provide a very high AC availability which is equated to a downtime of only 0.3 seconds/year.

IX. Fully Redundant Power System is one where every DC output of the power system is fault tolerant i.e. any fault in another parallel power system shall keep the system operational. Such a power system could be very expensive but this expense should be carefully evaluated in light of the importance of the service provided to the customer. If the system is fully redundant and has a reasonable MTBF of 500K hours, the total downtime per year could just be less than 1 second! However if the system is cost sensitive, a highly reliable power system could just very well be suitable for the application.

X. Other Important Considerations

1. **System Redundancy:** Even though a fully redundant power system provides very high availability, the overall system availability should be evaluated in terms of reliability of other functional blocks of the system such as Hard Disk Drives, I/O cards, memory and microprocessors. If any one of the system functional blocks has a higher failure rate than a DC/DC regulator in the power system, a fully redundant power system could not be of much help because the functional block with higher failure rate will fail more often than the DC/DC regulator.
2. **Current share :** Current sharing among redundant DC/DC regulators could be very problematic because of very high di/dt requirement. The output bus voltage might dip due to the slow current share response if one of the DC/DC regulator in parallel

very low ESR capacitors at the point of load to prevent voltage glitch.

3. **Efficiency:** An OR-ing MOSFET or diode used for fault isolation shall degrade the regulator efficiency by as much as 10% at very low voltages. However, if an OR-ing MOSFET is not used, failure of the low side MOSFET shall result in shutdown of the system.

XI. Centralized or Distributed Power

The general perception in the industry is that a Centralized Power Supply system costs less than a Distributed Power system. Such was the case when a system only required distribution of +12V, +5V, and +3.3V. The server systems of today and tomorrow shall operate at DC voltages below 1 volt. Even if the Centralized Power system could cost less, a Distributed Power Architecture is to be preferred for the following reasons;

1. **Transient Response:** A 100 A/usec or higher di/dt requirement is very hard to be met unless POL DC/DC regulators are used. The Centralized Power system shall not function properly due to path inductance and high di/dt unless very large number of low ESR capacitors are used at the point of load. In addition, the modern day processor requires currents of the order of 100A, and therefore distribution of such high currents drawn from a centralized power supply for multiple microprocessors on a system board is a big problem.
2. **Remote Sense:** Remote sense compensates for steady state voltage drop only. The parasitic path inductance and capacitance introduce higher order poles, and therefore control loop stability becomes an issue for applications requiring fast transient response time.

- 3 **Total Cost:** Even if the initial cost of a Centralized Custom Power system appears to be lower, the total cost could be much higher if the cost of development, quantities required and warranty costs are taken into account. A multiple output power supply is hard to design for products two generations away because of uncertainty of current requirement for multiple outputs from one product to another. It is much easier to design a single output AC/DC power supply for products which are two generations away thus allowing reuse of the existing power supply.
- 4 **Size:** A best designed system has more space allocation for system functions than for a power supply. A Centralized Power Supply system with an average power density of 3 W/cubic inch compared to an average of upto 10 W/cubic inch for a Distributed Architecture is thus clearly at a great disadvantage.
- 5 **Flexibility:** A Centralized Power Supply system is generally custom designed for a particular application and is rarely suitable for another application due to different power requirements for different systems. On the other hand, a Distributed Power Architecture could be configured for another application by reusing the existing AC/DC regulator and POL DC/DC regulators.
- 6 **Optimization:** At the start of any project, a Custom Designed power supply appears to be the best solution. This is not necessarily true because of uncertainty of power requirement for different system boards as the system is designed and modified. Every project starts with some power system requirements which are not very well defined at the start of a project, but by the time the actual hardware is qualified, the

the power needed is often found to be significantly different than the original estimated power. A Distributed Power system can be easily redesigned with minimum impact to the cost or the schedule in case the power requirements change.

XII. Distributed Bus Voltage:

Once a Distributed Power Architecture is selected based on its merits, the next big challenge for the power system architect is to decide what DC distribution bus voltage to use? The bus voltage could be as low as +5V DC and as high as 350V DC depending on the application. For example, if it is an IBM enterprise server system, the selected bus voltage is to be 350V DC because of the infrastructure already in place. On the other hand if it is a telecommunication application, -48V is used as the DC distribution bus voltage because of the battery plant voltage and the whole telecommunication infrastructure.

However, if the application is a server rack with multiple drawers using different machines from different vendors, the preferred method could be to provide AC to each drawer and then use an AC/DC regulator inside the drawer to provide +12V DC distribution bus. +12V DC distribution bus is selected because it provides 80% of the system load which includes input power to VRMs, Intel based microprocessors, Hard disk drives, fans and CDROMs. All future server system I/O cards shall require voltages below 2.5V and will need imbedded DC/DC regulators to be powered by +12V DC. A power system architect should therefore carefully evaluate the application and the future technology trend before deciding what DC distribution bus voltage to choose. If 350V DC bus is the best choice for one application, it might not be the best for another application. The most suitable bus voltage is one that provides maximum flexibility and could be suitable for future applications.

XIII Conclusion

The task of designing an optimal power system for a server system is very challenging. For a power supply vendor, understanding the specific application is very critical to recommend alternate power system architectures that are more suitable than what the customer has in mind. The paper covers various issues and trade off for designing a power system specifically for server applications. AC redundancy, AC/DC redundancy and DC/DC redundancy issues are described for the systems and an effort is made to highlight some very important issues concerning power system availability, imbedded or plug-in regulators, and Centralized Power systems vs. Distributed Power Architectures. The availability, the flexibility and the cost trade off are covered using specific examples. Ultimately, it is the power system designer's responsibility to have in-depth knowledge of the server system and future technology trends if a cost effective power system has to be designed.

Acknowledgment

I would like to thank Joseph Horzepa and Mohan Mankikar in providing valuable suggestions and reviewing the paper.