White Paper

How Green Is Your Cloud?

The broadband and smartphone revolutions have dramatically changed how we consume and interact with information in our professional and private lives. Storage and computing servers, colloquially known as the "cloud," increasingly host most of the major applications. They use fast internet connections and extremely powerful resources to consolidate and process far-flung data and respond rapidly to users. This paradigm holds the promise of instant and always-on response, ubiquitous access and significantly lower capital investment for clients. Cloud servers are an indispensable part of our daily activities whether it's for consumer applications like Netflix, Facebook and Siri, or the industrial Machine-to-Machine (M2M) activity that is the basis of the Internet of Things (IoT), or enterprise solutions such as SAP and salesforce.com.

Growth of the Cloud

Driven by these forces, server capacity has grown at a remarkable rate. Cloud computing is estimated to have grown from essentially zero in 2006 – when Amazon launched its Amazon Web Services (AWS) platform – to a $58 billion market in 2014; and according to Forrester Research, the public cloud market (excluding captive datacenters from Amazon, Google and others) is predicted to reach $191 billion by the end of the decade. By way of comparison, the WSTS forecasts the entire semiconductor industry to reach $333 billion in 2014 worldwide sales. Figure 1 shows the capital expenditure at Amazon, primarily driven by computing horsepower in its datacenters.

Figure 1. As cloud services have grown, companies like Amazon have exponentially increased spending on computing equipment.

Cloud computing revenues include the effects of lower prices as the technology matures and competition increases. Even more amazing has been the cloud's growth in raw compute horsepower. By one estimate, AWS, the leading cloud service provider, has deployed more than 2.8 *million* servers worldwide.

Architecturally, servers have also evolved significantly by moving to hyperscale and multi-threaded structures, and processor cores have significantly improved their raw throughput. Additional design techniques, such as being able to change clock speed, supply voltage on the fly and vary the number of simultaneously operational cores, have all enabled greater dynamic response to computing load requirements. Nevertheless, they have also added significant complexity to power delivery requirements.

Growing Consumption

Most relevant is the electrical energy consumed by these datacenters: as the installed base of servers has mushroomed, so has their energy consumption. Obtaining exact company numbers is difficult, but one exemplary datacenter designed to consume 3MW of power hosts more than 8000 servers. Google estimated in 2011 that its datacenters alone continuously draw about 260MW, which is about 25% of the output of a state-of-the-art nuclear power plant. To ease power transmission challenges, the vast majority of the datacenters worldwide are located close to massive power sources like the Columbia River hydroelectric power system. Worldwide 2010 estimates stated that datacenters used between 1% and 1.5% of global electricity consumption, which is equivalent to Brazil's total consumption. In the U.S., datacenters use closer to 2% of the country's total electricity consumed, which is the equivalent of New Jersey's total consumption. Essentially, datacenters had added another New Jersey to the U.S. electric grid by the end of 2010, and the load continues to grow.

This massive power consumption growth has a significant economic impact. While processor cores may offer greater processing capability, as they follow Moore's Law and architecture improvements, their voltages have not scaled fast enough to lower overall power consumption. Datacenters primarily use power in two ways: to supply the needed energy for the computers and secondly to cool them sufficiently, keeping systems within their operational range. Consequently, small improvements in the efficient delivery of power have a leveraged beneficial impact on the bottom line. In addition to the reduced power bill, efficient power delivery can also increase datacenter capacity for a given budget – a very important consideration given that the installed capacity is continuing its brisk pace of double-digit annual growth.

Power Distribution

State-of-the-art datacenter power distribution consists of a series of stepped-down voltages followed by point-of-load power delivery. Raw efficiency is the greatest challenge, but power systems can also add capabilities in a few different ways to enable lower energy consumption.
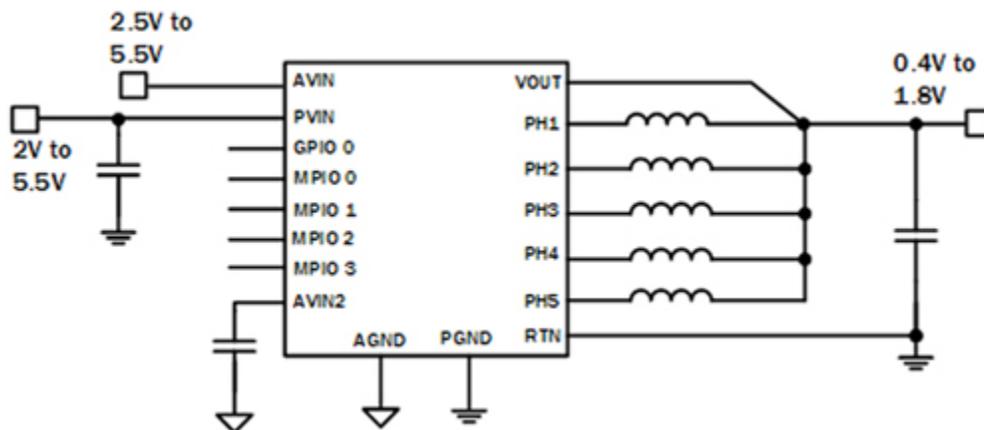

Figure 2: A typical multi-phase power system

Multi-phase Operation: The latest generation of infrastructure power converters supports multi-phase operation, and maintains the power delivered from light-load to peak-load at a close to peak efficiency. They achieve close to peak efficiency by parallelizing multiple power delivery phases, which the controller modulates based on the power draw requirement. Figure 2 shows a typical multi-phase system, where the regulator modulates each inductor to provide a variable amount of current. Multi-phase operation improves on a key shortcoming of single-phase converters, where the efficiency peaks at a nominal load, but drops at very high loads. Figure 3 shows how a multi-phase system can intelligently select the number of phases depending on the load. Flattening the efficiency curve across a much larger portion of the operating range frees datacenter planners from choosing between optimizing for typical and maximum workloads.
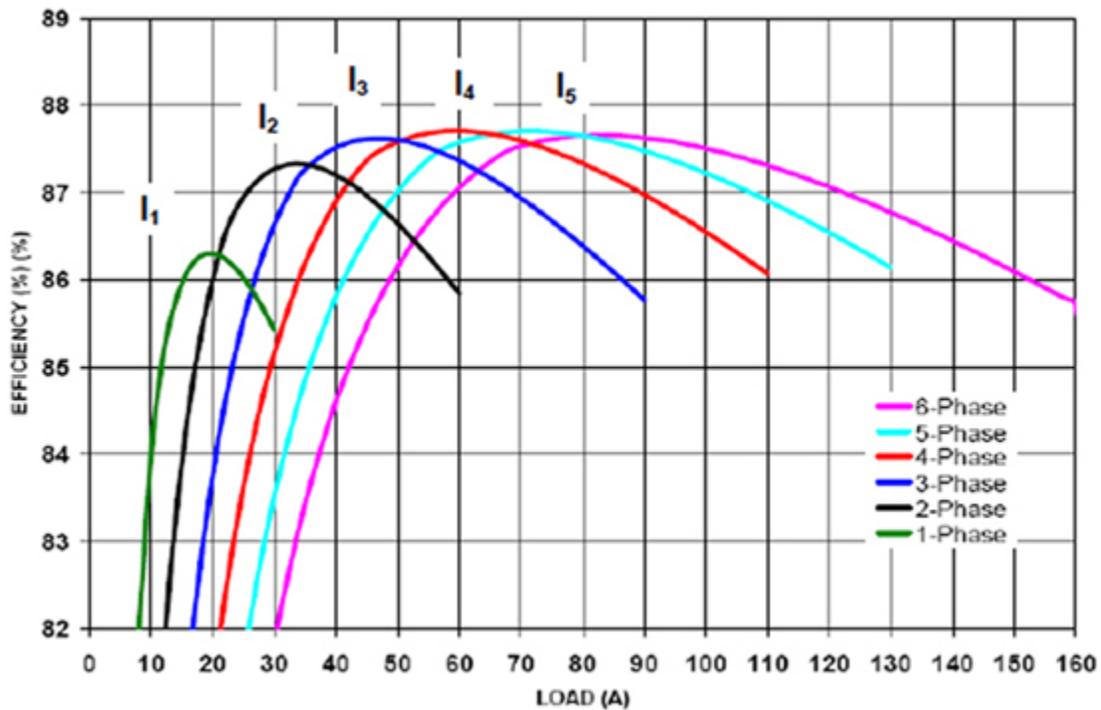

Figure 3: Multi-phase power regulators maximize power efficiency across a wide operating range

Parallel Power Channels: Dividing processor cores into "power islands" enables parts of the system to shut down when they are not in use. Power delivery systems now comply with the need to provide multiple simultaneous rails.

Communication: Today's processor cores communicate anticipated power capacity to the power converters via a digital bus (typically PMBus). The changing load can be a function of additional cores coming on-line, variation of processor clock speed, or the knowledge that the software is processing a particularly intensive sequence. With insight into the expected loading, the controllers are able to maximize the efficiency across the load curve. The ability to manage the duration and level of energy consumed provides another big advantage to service providers: they can use the duration and intensity of system activity to calculate the billings for each process.

Green Lining

There is a green lining to this "cloud." Even if it uses a significant amount of energy, the efficient consolidation of computing tasks in the cloud – which employs some of the most powerful computation

systems ever to exist – still holds the promise of lowering the overall amount of energy required to perform tasks across discrete systems. The datacenter's evolving power delivery requirements offer semiconductor manufacturers significant opportunities to innovate and create new breakthroughs.

Intersil offers best-in-class power products for infrastructure applications, including multi-phase controllers, advanced power stages and point-of-load regulators. For more information on Intersil's power management solutions, please visit www.renesas.com/power.

References

1. Gartner Research Report: Amazon's Cloud Revenue is Important Information for Potential Customers, January 30, 2013.

Next Steps

- Learn more about Renesas' power solutions for infrastructure applications
- Search for power management parts using our parametric search

# # #