White Paper

Intel® Processors for Web Architectures

Optimizing Web Infrastructure on Intel® Architecture

Executive Summary and Purpose of this Paper

Today's data center infrastructures must adapt to mobile usage models and cloud service challenges while at the same time scaling to meet escalating "big data" traffic demands. Now more than ever, data center managers see the administrative benefits of maintaining dedicated server tiers for specific tasks such as serving Web pages, executing application services, and hosting databases.

This paper focuses on optimizing the design and deployment of Web-tier services using Intel® based servers, communications, and storage solutions to achieve the best balance of high-performance, low power consumption, high-availability, affordability and low Total Cost of Ownership (TCO).

Table of Contents

The Challenges of Web-Tier Design

An effective Web-tier needs to address a variety of sometimes conflicting objectives, including maximizing performance for every end-user, with potentially millions of users being managed at the same time. In addition to consistently meet performance levels, the Web-tier needs to incorporate a scalability strategy to handle rapidly escalating volumes of data and numbers of new users, without re-engineering the basic architecture along the way.

Of course, cost is a key consideration; not just the capital expenditures for the hardware and software, but also the operating costs. These include administrative personnel and the largest expenditures for operating many data centers today, electricity and cooling. According to a survey conducted in 2012 by the U.S. Uptime Institute, more than 40 percent of data center managers expect to run out of power capacity within 12-24 months.

As described in this paper, Web-tier design considerations vary depending on business requirements. For example, a cloud giant such as Google or Facebook runs different workloads and delivers services to a different size and type of user base than typical Enterprise Data Centers do. But there are many commonalties as well. A variety of specialized Web-tier architectural approaches and microserver solutions emerging in the Enterprise today were pioneered by giants such as Google, Yahoo!, and Facebook together with the open source community.

The following sections answer these key questions about building your Web-tier infrastructure:

- What resources are required to meet requirements for performance, security and scalability?
- When should you scale-up on a single server, or scale-out with many servers?
- How should you optimize your network communications?
- What type of storage solution fits best?
- How do the different usage models of cloud service providers vs. Enterprise data centers affect these considerations

Web Infrastructure Considerations

Web servers inherently differ in deployment approaches and support needs compared to servers in other tiers. While mid-tier application and back-end database servers require fault-tolerance and high capacity utilization to efficiently handle business transactions, front-end Web servers must deliver a highly responsive end-user experience. That requires rapidly handling user requests for content and services residing in other tiers.

Security requirements also may differ between the Web-tier and other tiers. In many enterprise infrastructures, the Web-tier sits within a "demilitarized zone" (DMZ) outside the firewall. In this case the web server takes responsibility for authenticating all users through https requests over SSL. When a web server sits inside the firewall, it can make use of an Application Delivery Controller (ADC) which handles user authorization and can offload all SSL web requests between the Web-tier and the firewall.

Web Server Resource Requirements

Today's enterprise Web tier may deploy thousands of server instances in a user-facing array that, depending on the business type and usage model, may need to scale to millions of simultaneous user interactions.

The ability to parallelize tasks and the relatively small users/Web server ratio makes the "scale-out" approach quite appropriate for the Web tier.

With that scaled-out architecture, the Web tier tends to consume more network sockets and I/O throughput than other tiers.

It's Not a One-Size-Fits-All Approach

Web server workload and I/O requirements vary significantly depending on whether the content is primarily static or dynamic. Traditional "Web 1.0" static content is typically stored as HTML pages. In contrast, "Web 2.0" dynamic content is stored in databases and changes with input generated by users or by software.

The Web 2.0 model enables applications such as online credit card transactions and social media interaction; however, it requires more and faster interaction with the application and database tiers.

For Web 2.0 services, I/O demands often require special caching layers, such as a Flash tier or a cluster of memcached servers, to provide fast access to hot-key information like trending social media topics.

Cloud Giants vs. Enterprises

Web tier architectures differ for cloud giants like Facebook and Google in comparison to Enterprise Data Centers. The cloud giants need to serve very large volumes of end-user requests at the lowest possible cost. That makes response rates and energy consumption the most critical considerations.

Enterprise data centers typically face more complex challenges because they run a variety of mission-critical and business-critical workloads. They need highly-reliable and secure environments, and they tend to control costs by virtualizing servers for efficient capacity utilization.

Those differences impact data center architecture decisions, particularly in scaling the Web-tier.

Web Server Scaling Methods

Web tier server deployments can scale in a number of different ways, including scaling "out", "up" and "both".

Scale Out

"Scale-out" means adding server capacity by simply adding more servers. The model works well for highly parallelizable workloads. In a scale-out infrastructure, each server has its own memory and often directly attached storage, as well. This hyper-scale, shared-nothing approach (also called "physicalization") offers the advantage of simplicity. A base server configuration can be replicated as many times as required.

Depending on specific needs, scale-out configurations can be optimized with Intel® Xeon® E3 processor-, Xeon E5

processor- or Intel Atom® SoC-based systems.

For many scale-out workloads, servers based on Intel Xeon E3 processors can offer the best combination of cost, power and performance. Higher compute demands can make dual-socket Xeon E5 processor-based servers the better choice. Atom SoC-based microservers with Thermal Design Power (TDP) levels as low as 6W, offer an excellent scale-out alternative for computationally light but I/O-intensive workloads.

Scale Up

At the other end of the spectrum, most enterprise database servers use a scale-up approach based on virtualization to optimize efficiency. Scale-up means adding server capacity by populating a symmetric multi-processing (SMP) or massively parallel processing (MMP) architecture with more processors, memory and storage.

That enables numerous virtual machines (VMs) to run workloads on a single large-scale physical machine. Intel Xeon E7 processors with up to 10 cores and 20 threads, 30 MB of on-die cache and up to 4096 GB of memory capacity, offer an excellent foundation for building massive scale-up processing environments.

Although large-scale SMP and MPP systems rarely run the Web-tier, several leading vendors offer Xeon processor-based Enterprise Data Warehouse appliances with hundreds of cores and terabytes of memory, able to process petabytes of data.

Solutions built on a single, large-scale system require greater Reliability, Availability and Serviceability (RAS) capabilities to mitigate the risks of a single point of failure. Self-monitoring and self-healing capabilities at the server level help to assure continuous health and optimal performance of the entire virtualized infrastructure. The Intel Xeon E5 and E7 processor families include built-in RAS features. They support active and passive management of server interconnects, data stores, data paths, and other resources, with the ability to proactively and reactively repair errors and avoid future problems.

Scale Up and Out

A scale-up-and-out approach entails first scaling up to the limit on each virtualized SMP system to support as many VMs as possible, and then scaling out by adding another SMP system. The up-and-out approach maximizes both scalability and fail-over availability for workloads with significant compute and I/O requirements.

Intel Xeon E5 processor-based 2-socket servers account for the majority of all systems in today's virtualized data centers and cloud environments. With up to 80% higher performance than their predecessors, Xeon E5 4600, 2600, 2400, and 1600 product families offer more cores, cache and memory capacity, along

with faster communication pathways to move data more quickly.

No matter what combination of Intel processors run the servers in your data center, the x86 instruction set enables one binary code base across all your workloads.

From a security stand-point, that means you get data center-wide support for Open SSL using built-in Intel Advanced Encryption Standard instructions. From a broader perspective, the single code base lowers TCO by simplifying your IT operations.

Scaling Approach	Key Issues	Processor Choices
Scale-Out	• Physicalization • Shared-nothing approach • Simplicity for scaling • Power conservation	• Intel Atom SoC • Intel Xeon E3 processor family • Intel Xeon E5 processor family
Scale-Up	• Virtualization • Optimize overall efficiency • Shared resources • Security, management and overall system health	• Intel Xeon E5 processor family • Intel Xeon E7 processor family
Scale-Up-and-Out	• Virtualization • Optimize overall efficiency • Shared resources • Security, management and overall system health	• Intel Xeon E5 processor family • Intel Xeon E7 processor family

Optimizing Communications

Web tier deployments inherently involve I/O-intensive operations. Insufficient network bandwidth and inadequate CPU-to-Network Interface Card (NIC) performance can cause latency issues that degrade the end-users' experience. More than anywhere else in your data center, the Web-tier requires high-bandwidth communications.

Bandwidth Choices: From 1G to 100G

Web server infrastructures commonly use 1Gbps Ethernet (1GbE) for server-to-server communications and often for communication with other tiers. But more data centers will deploy 10GbE on Web servers as the cost continues to decrease and as demands continue to grow.

The main consideration for achieving optimal cost/performance is to match the NIC with the processor's compute capacity.

For instance, most hyper-scale scenarios using ultra low power Atom SoC-based microservers achieve the right balance by using1GbE.

On the other hand, a scale-out or scale-up-and-out deployment using more powerful Intel Xeon processors typically would require 10Gbps communications to take full advantage of the high compute capacity of the servers.

For large enterprise data centers, communications between the Web tier and other tiers most often uses 10GbE or 40GbE, depending on specific requirements.

In the not-too-distant future,100Gbps silicon photonics technology will drive high-speed rack-scale communications in data centers. Intel recently demonstrated 100Gbps solutions in rack-level scenarios, opening up a new wave of hyper-scalability by removing a traditional data center communications bottleneck.

I/O Virtualization

The optimization of Web-tier networking also benefits from I/O virtualization. For example, Intel VT-c can optimize data paths between VM instances and the NIC while also managing network Quality of Service (QoS) to ensure that each packet arrives at its destination once and only once.

Bypassing hypervisor software to allow direct NIC hardware access by VMs reduces CPU overhead, minimizes latency and increases network throughput.

Tuning the NIC queue-affinity via script-level control of factors such as multi-queue handling, interrupts, thread migration control, and load balancing between cores also helps optimize the usable bandwidth.

Communications Level	Key Issues	Product Solutions
Intra-tier clustered server level	• Proper NIC-to-Processor load matching • Power/cost/performance	• Intel 1GbE • Intel 10GbE • Intel 10 GbE with VT-c
Tier-to-Tier optimization and Data Center level	• Rack-level aggregation • High-speed switching	• Intel 10 GbE • 40 GbE • Future Intel Silicon Photonics 100Gbps Solution

Storage Considerations

Storage architectures for Web tiers can be designed with Direct Attached Storage (DAS), Network Attached Storage (NAS) or Storage Area Network (SAN) approaches.

The DAS approach works well for local file sharing with just a few servers. It also is ideal in hyper-scaled physicalization models where each server (or each processor core) needs its own storage.

Solid State Drives (SSDs) speed-up read/write 100x faster than Hard Disk Drives (HDDs) in a DAS configuration. Intel SSD 710 Series SATA SSDs combine the low power, performance and reliability needed for DAS deployment and scaling within the Web tier.

NAS works well for file-level data sharing such as in a Web 1.0 environment where clusters of servers handle large numbers of user requests for static content. Dedicated NAS devices can avoid data duplication and offload storage requirements from servers.

SANs provide high-availability for moving large blocks of data such as images or video. SANs frequently connect to servers through iSCSI, and Intel Ethernet Server Adapters provide a simple, reliable and cost-effective method to connect servers to iSCSI SANs without the need for special hardware or software.

Storage efficiency techniques such as data compression, data de-duplication, intelligent tiering and thin provisioning also help to optimize data handling in Web tiers. For example, intelligent tiering enables "hot" data to move automatically to faster storage, such as SSDs, thereby giving Web servers faster access to trending data.

Storage Approach	Applicability	Product Solutions
Direct Attached Storage	• Physicalized server clusters • Scale-out approaches	• Intel SSD 710 Series
Network Attached Storage	• Small cluster file-sharing • Larger clusters with static data • Enterprise data centers	• Intel Ethernet Server Adapters
Storage Area Network	• Large block data (image, video) • Enterprise data centers	• Intel Ethernet Server Adapters

Web-Tier Optimization

As previously mentioned, Web tier requirements differ for Cloud Giants as compared to Enterprise Data Centers. Although many issues outlined in this paper apply to both situations, it is useful to focus specifically on the needs of each.

Cloud Giants

For companies like Facebook and Google, the data center represents their "factory". They want to maximize performance while minimizing the cost of operations.

The biggest cost factor for the cloud giants is not CapEx, staffing, software or maintenance, but the electricity needed to run the data center and cooling systems.

Workloads focus on serving large volumes of end-user requests in parallel, using physicalized Web tiers (sometimes with many thousands of servers per data center). Such I/O intensive workloads are best addressed by microservers based on Xeon E3 processors or Atom SoCs. The Atom-based approach can be particularly useful for power conservation with as little as 6W Thermal Dissipation Power (TDP) per SoC. Intel Ethernet technology matches to optimize CPU loading, with 10GbE best for Xeon- and 1GbE best for most Atom-based systems.

Enterprise Data Centers

Enterprise data centers face more complex challenges because of their mission-critical and business-critical tasks.

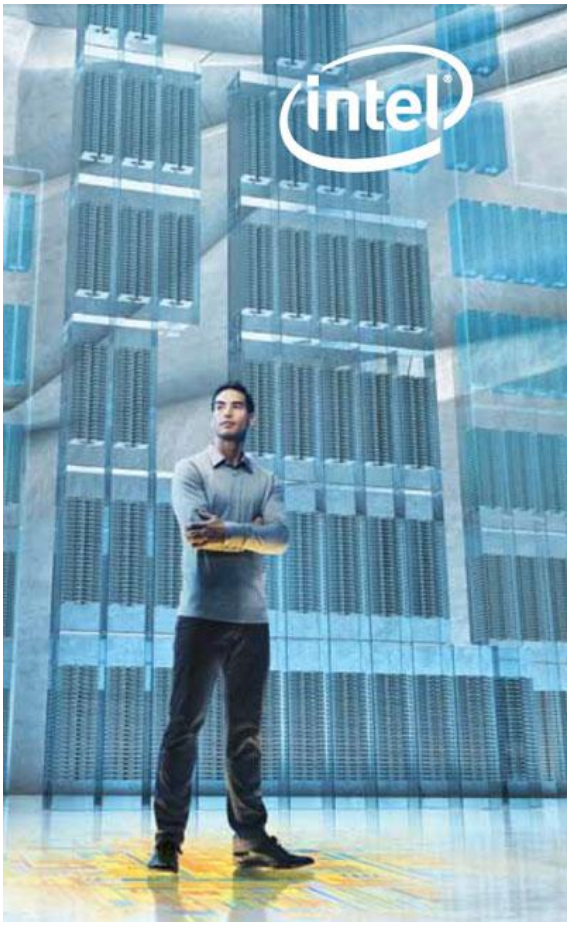
Although enterprise IT departments care about power consumption too, other important considerations focus on very high levels of information security, system availability, and efficient utilization of server assets. Therefore, enterprise data centers tend toward a scale-up-and-out model that emphasizes virtualization of as many VM instances as possible per server before scaling out by adding more servers.

Xeon E5- or E7-based servers and 10GbE meet these requirements best, in combination with virtualization tools like Intel VT-c.

Web Tier Type	Challenges	Solutions
Cloud Service Providers (CSPs)	• Maximize physicalization • Minimize power usage • Scale I/O intensive and easily parallelized loads	• Intel Atom SoC-based microservers with 1GbE • Intel Xeon E3 or E5 processor-based servers with 10GbE • Intel SSDs
Enterprise Data Centers	• Scale variety of complex mission-critical loads • Maximize efficiency with virtualization • Maximize information security	• Intel Xeon E5 or E7 processor-based servers with 10GbE • Hardware acceleration for virtualization, such as Intel VT-c

Summary

Using a tiered architecture helps optimize overall data center efficiency, security and scalability. Separating the front-end Web tier from the application tier and from back-end database tier enables IT managers to focus on the differing requirements performance, scalability, information security and systems management requirements for different workloads.



Although requirements differ across different types of data centers, every Web tier implementation needs to optimal scalability, reliability and power efficiency.

Whether your particular strategy entails scaling out with clusters of many physical servers or scaling up by using virtualization to maximize efficient utilization of every server using multiple VMs, the broad ecosystem of Intel-based processors, software, communications, and storage solutions can be leveraged to deliver optimal performance, power usage and future scalability.

For more information about Intel server technologies and products, visit **the Intel Server Page**

For more information about Intel microserver technologies, visit **the Microservers Powered by Intel Page**

For more information about Intel storage technologies and products, visit **the Intel Storage Products Page**

For more information about Intel network technologies and products, visit **the Intel Ethernet Controllers Page**

Intel® Virtualization Technology requires a computer system with an enabled Intel® processor, BIOS, and virtual machine monitor (VMM). Functionality, performance or other benefits will vary depending on hardware and software configurations. Software applications may not be compatible with all operating systems. Consult your PC manufacturer. For more information, visit http://www.intel.com/go/virtualization

This document contains information on products in the design phase of development.

Intel processor numbers are not a measure of performance. Processor numbers differentiate features within each processor family, not across different processor families: Go to: Learn About Intel® Processor Numbers

No computer system can provide absolute security under all conditions. Built-in security features available on select Intel® Solid State Drives may require additional software, hardware, services and/or an Internet connection. Results may vary depending upon configuration. Consult your system manufacturer for more details.