AMD ROME-OPTIMIZED PLATFORMS

Selecting a new Server Platform That is Workload-Optimized with the 2nd Generation AMD EPYC Processor with up to 64 Cores

An IDC InfoDoc, sponsored by Dell EMC | August 2019

Analyst: Ashish Nadkarni
Enabling Business Outcomes at the Speed of Innovation

Executive Summary

We are noticing a marked shift in the digital economy in terms of how businesses rely on infrastructure. Business priorities are driving workload transformation and placement, which in turn are driving infrastructure transformation. Investing in “right-fit” workload infrastructure is one of the crucial pillars of this transformational journey. AMD 2nd Generation EPYC processors with up to 64 cores (codenamed “Rome”) represent a unique opportunity for addressing the modern workload-optimized infrastructure world. However, identifying the server platform that takes advantage of this opportunity is not always straightforward. Some server platforms may support the new processor, but don’t enable the new processor features. Businesses must check the processor specs against the vendor’s server specs and ensure the server platform is a fully optimized Rome platform. This InfoDoc will help guide you in the process and help you confirm long-term return on investment and reduced total cost of ownership.

To support superior business outcomes, CIOs must make strategic investments in IT infrastructure to support current and next-generation workloads. This infrastructure must:

• Be highly available and scalable
• Support end-to-end security
• Support high performance compute
• Enable fast data access

For the CIO, top business priorities for IT infrastructure are

• Reduced TCO – Total cost of ownership
• Superior ROI – Return on Investment
• Elasticity
• On-demand Performance
• Security and Compliance

* The term “Workloads” is a way of describing applications and their existing datasets. Certain types of workloads are used to hold certain types of datasets e.g., SQL databases for structured “schema-conforming” datasets and NoSQL databases for “schema-less” semi-structured datasets.
Choosing a Secure, Right-Fit Infrastructure Platform for Your Most Critical Workloads

A “right-fit” infrastructure platform incorporates technology innovations that address critical business priorities:

- Dynamic workload-optimized scaling
- Effortless management
- High bandwidth capacity
- Low latency data access
- Faster network connectivity
- Robust end-to-end security

Why should “right-fit” infrastructure matter to you?

For the CXO
- Maximized return on investment
- Reduced total cost of ownership
- Security of data and intellectual property
- Backed by the capabilities of a trusted vendor

For the Enterprise Architect
- Optimized scaling
- Faster hardware
- Effortless systems management
- Larger bandwidth capacity
- Robust built-in security that starts at the processor

A server platform provides a solid foundation on which firms launch their digital initiatives. The security of this server platform is crucial to maintaining the integrity of the data that lives on it. Therefore, server security is a crucial starting point when making any hardware purchasing decisions. Security must be core to the design of the platform and not be “bolted on” as an afterthought. A recent study found that when evaluating server platforms, respondents ranked hardware security (including BIOS, BMC, and other firmware) higher in importance than total cost of operations.

Security Ranks Number 1 When It Comes to Evaluating the Purchase of Server Infrastructure (top 6 showing)

- Security of hardware: 41
- Total cost of operations: 34
- Efficient provisioning: 30
- Improved staff productivity: 29
- Capital costs: 28
- Efficient management: 27
One Size Does Not Fit All: Why the Workload Drives the Selection of a Right-Fit Infrastructure Platform...

A modern business is a technology and data business. It has multiple functions that need to be managed efficiently and executed in a timely manner. Be it simple functions such as file and print, customer relationship management applications, or applications that use artificial intelligence for predictive outcomes.

Each business has unique workload requirements – performance, agility, or investment returns – which in turn drive the mix of current and next-generation workloads. Multicloud environments (which are becoming the new normal) place unique demands on IT infrastructure. The IT infrastructure for each business must support any of these operational or strategic needs. The selection of an infrastructure platform (servers) therefore must be performed in a workload-centric manner – along four dimensions:

The Four Dimensions of An Infrastructure Platform

Performance: How fast can my platform run? How can I accelerate the performance?
Compute Scalability: How well can my server systems scale for compute? An easy way to think about this is on sockets/U-height basis.
Acquisition Cost: What kind of choices are available on a cost/node basis?
Storage Capacity: What kind of storage capacity is available on the system? This is a critical factor for several data-intensive, low-latency applications and this is expressed in TB/node.
...And Then Shifts to Taking a Closer Look at the Profile of Key Business-Critical Workloads

The approach outlined above must be the first step in a workload-optimized infrastructure platform selection process. It is imperative for enterprise systems architects to examine each workload type for its requirements on compute, I/O, connectivity, and storage optimization. For example:

- **Data analytics** workloads require matching compute
- **HPC** workloads need bandwidth and memory
- **AI/ML/DL** workloads present IT architects with a key challenge in choosing a platform with the right performance characteristics to tackle complex compute-intensive algorithms
- **Desktop virtualization** workloads demand reliable performance and bandwidth to match dynamic user demand
- **In-memory database workloads** require that the platform support the bandwidth and throughput to manage high-volume transactions

### Compute and Storage Workload Profile

<table>
<thead>
<tr>
<th>Data Analytics and AI/ML/DL</th>
<th>Software-defined Storage</th>
<th>Massively-Parallel Computing</th>
<th>Virtualization and VDI</th>
<th>Network Virtualization</th>
</tr>
</thead>
<tbody>
<tr>
<td>High parallelism for complex analysis</td>
<td>Direct high-performance and capacity-optimized storage support</td>
<td>Highly parallelized processor optimized for accelerators and real-time data streams</td>
<td>Single Socket performance that reduces TCO without compromising availability</td>
<td>Accelerated Provisioning/Agility</td>
</tr>
<tr>
<td>Massive I/O bandwidth for faster data loads</td>
<td>Low latency and high I/O parallelism for data access and persistence</td>
<td>Massive I/O for cluster connectivity</td>
<td>Higher core count to enable dense user base</td>
<td>High bandwidth network connectivity</td>
</tr>
<tr>
<td>High-integrity for floating point capacity</td>
<td>More memory for larger cache</td>
<td>Memory capacity for larger datasets</td>
<td>Larger memory capacity for VM density</td>
<td>High integer and floating-point capacity</td>
</tr>
<tr>
<td>Security for business-critical data</td>
<td>Memory encryption for data security</td>
<td>Massive I/O bandwidth</td>
<td>Cryptographic isolation between hypervisor and VM</td>
<td>Security for business-critical data</td>
</tr>
<tr>
<td>Hadoop/Spark</td>
<td>Hyperconverged and non-hyperconverged file, block and object SDS</td>
<td>High-performance applications</td>
<td>VMware Horizon, Citrix Xen Desk/Xen App, Autodesk</td>
<td>OpenStack, NSX, AHCI</td>
</tr>
<tr>
<td>Compute optimized (with support for GPUs/accelerators)</td>
<td>Storage and I/O optimized</td>
<td>Compute optimized</td>
<td>Compute and I/O optimized</td>
<td>Compute, I/O, and connectivity optimized</td>
</tr>
</tbody>
</table>
**Demanding and Unique Workload Profiles Drive the Choice of an IT Infrastructure Platform**

The selection of a suitable infrastructure platform (i.e. server) must start with the processor and processor-centric optimizations (e.g., memory, I/O, connectivity and storage)

<table>
<thead>
<tr>
<th>Attribute Category</th>
<th>What to look for</th>
</tr>
</thead>
<tbody>
<tr>
<td>Processor Sub-system Performance</td>
<td>• High Core and thread count per socket</td>
</tr>
<tr>
<td></td>
<td>• High L2 and L3 cache per core</td>
</tr>
<tr>
<td></td>
<td>• Higher TDP wattage</td>
</tr>
<tr>
<td></td>
<td>• Memory Performance and Capacity</td>
</tr>
<tr>
<td>I/O Bandwidth and Latency</td>
<td>• PCIe capabilities for high bandwidth/low latency access to accelerators,</td>
</tr>
<tr>
<td></td>
<td>connectivity interfaces and storage media</td>
</tr>
<tr>
<td></td>
<td>• Number of PCIe lanes – flexibility in configuration for optimized performance</td>
</tr>
<tr>
<td>Storage Performance and Latency</td>
<td>• NVMe direct performance</td>
</tr>
<tr>
<td></td>
<td>• PCIe lanes for non-direct I/O</td>
</tr>
<tr>
<td>Processor Sub-system Security</td>
<td>• Processor security to prevent hostile code execution attacks</td>
</tr>
<tr>
<td></td>
<td>• Memory encryption</td>
</tr>
<tr>
<td></td>
<td>• Workload and VM-aware isolation</td>
</tr>
<tr>
<td></td>
<td>• Key management</td>
</tr>
<tr>
<td></td>
<td>• Operating system and hypervisor support</td>
</tr>
<tr>
<td>Operating and Ownership Costs</td>
<td>• x86-based ISA eliminates the need to rewrite applications</td>
</tr>
<tr>
<td></td>
<td>• Single-socket performance for reduced TCO in socket-based licensing</td>
</tr>
<tr>
<td></td>
<td>• Dual-socket performance for improved ROI in mixed or specialized workload</td>
</tr>
<tr>
<td></td>
<td>configuration</td>
</tr>
<tr>
<td></td>
<td>• Based on industry-standard x86 ISA requiring no migration costs</td>
</tr>
<tr>
<td></td>
<td>• Integration with existing management solution</td>
</tr>
</tbody>
</table>

The processor is a crucial building block for an infrastructure platform. The choice of a suitable platform must start with a critical examination of the processor sub-system. Many times, this is taken for granted, and can have serious repercussions on business outcomes when workload performance or scaling suffers due to lack of capabilities, availability, or resources. Or, in many cases, the infrastructure becomes vulnerable due to processor design shortcomings.
The AMD 2nd Generation EPYC Processor with up to 64 Cores Provides a Critical Building Block for an Optimized, Scalable and Secure IT Infrastructure

Advanced Micro Devices (AMD) introduced its AMD EPYC brand of x86-64 microprocessors in 2017. Based on the company’s Zen microarchitecture, the server variant of these processors are specifically targeted for computationally and I/O-intensive datacenter and cloud workloads. With features such as higher core counts, more PCIe lanes, support for larger amounts of RAM, and larger cache memory, these processors support single and dual-socket system configurations. Naples is AMD’s first generation EPYC server processor family followed by Rome.

<table>
<thead>
<tr>
<th>Workload Attribute</th>
<th>AMD EPYC capabilities</th>
<th>Naples</th>
<th>Rome</th>
</tr>
</thead>
<tbody>
<tr>
<td>High compute parallelism</td>
<td>Higher core count with configurable TDP</td>
<td>• 32 cores</td>
<td>• 64 cores</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• 64 threads</td>
<td>• 128 threads</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• 2 x L3 cache/core (16MB per 4 cores)</td>
<td>• 2 x L3 cache/core (16MB per 4 cores)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• 200W TDP</td>
<td>• Up to 240W</td>
</tr>
<tr>
<td>Faster in-memory workloads</td>
<td>RAM capacity and speed</td>
<td>• 2666 MT/s</td>
<td>• 3200 MT/s</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Max. system memory 18TB</td>
<td>• Max. system memory 32TB</td>
</tr>
<tr>
<td>Massive I/O for data access and connectivity</td>
<td>More PCI lanes</td>
<td>• Up to 128 PCI Gen 3 lanes xGMI 10GT/s</td>
<td>• Up to 160 PCIe Gen4 lanes xGMI-2 at 16 GT/s</td>
</tr>
<tr>
<td>Cryptographic isolation between hypervisor and VM</td>
<td>Secure Encryption Virtualization (SEV)</td>
<td>• Up to 16 SEV keys per hypervisor</td>
<td>• Up to 509 SEV keys per hypervisor</td>
</tr>
<tr>
<td>Security for business-critical data</td>
<td>Secure Memory Encryption (SME)</td>
<td>• Not supported</td>
<td>• Configurable</td>
</tr>
</tbody>
</table>

Advantages of the next-gen EPYC (Rome) processor platform

- Single I/O/memory die removes processor subsystem bottleneck, and provides 2x performance over Naples
- Uniform speed across cores, and high bandwidth and low latency network connectivity
- Separate I/O die from Zen2 chiplets* allows flexible core configurations
- Dedicated encryption offload sub-processor provides full stack encryption without any overhead
- Dedicated I/O plus memory die provides improvements for NUMA** workloads while also reducing latency for non-NUMA workloads
- Improved single socket and reimagined dual socket performance reduces TCO

* Chiplets are multiple smaller pieces of silicon that make up a processor vs. the older process of carving new processors from silicon as single chips
** Non-uniform memory access (NUMA) is a computer memory design used in multiprocessing, where the memory access time depends on the memory location relative to the processor. Under NUMA, a processor can access its own local memory faster than non-local memory (memory local to another processor or memory shared between processors). The benefits of NUMA are limited to workloads where the data is often associated strongly with certain tasks or users.
What to Look For in a “Rome-Optimized” Server Platform

A "Rome-Optimized" server platform is one that has been designed from the ground up to take full advantage of capabilities offered by the 2nd generation AMD EPYC Processor with up to 64 cores. Such server platforms offer investment protection as their capabilities are future-ready.

Did you know? Just because the features are on the AMD processor doesn’t mean the server has those features. Each server manufacturer decides which features and capabilities to use. Always check the server specs to ensure you are getting an AMD Rome-Optimized platform.

### Design capabilities that enable optimization

<table>
<thead>
<tr>
<th>Rome-Optimized Platforms</th>
<th>Design capabilities that enable optimization</th>
</tr>
</thead>
</table>
| Are built with a high-performance processor sub-system, reduce latency, and deliver fast response times | • Support for full processor stack – up to 240W TDP  
• 20% faster memory speed at 3200MT/s |
| Enable faster data access, transfer speeds for storage workloads, and faster connectivity for networking and storage workloads | • 25% more PCIe lanes and 2x PCIe speed - up to 160 PCIe Gen4  
• 60% faster interconnect fabric with xGMI-2 at 16GT/s (Up to 24 GT/s with a supported peripheral) |
| Are designed to make full use of processor capabilities for integrated end-to-end security | • Secure data with AMD Secure Memory Encryption (SME)  
• Workload VM to hypervisor isolation with AMD Secure Encrypted Virtualization (SEV) - 509 unique keys per hypervisor |

### Additional configuration enabled by vendor’s design and engineering

- Design makes use of higher sustained TDP of 240W per processor
- Risers optimized to balance workloads (flexible configuration)
- Flexible use of PCIe lanes for network and management access
- Socket direct networking for balanced network I/O
- Direct NVMe configuration eliminates switches or bridges
- Extended speed mode ready for faster peripherals operating at 24 GT/s when available
- BIOS supported AMD Secure Memory Encryption
- Liquid cooling options via partners
- Reduction in deployment time through automation capabilities and system management solutions
- Improvement in datacenter cooling power utilization efficiency (cooling PUE)
- Easy BIOS tuning with workload-optimized server configuration profiles
- Verifiable secure root of trust that is etched at the factory
A Trusted Vendor Delivers a Best-in-Class “Rome-Optimized” Server Platform

Organizations must consider the role of IT vendors as partners when selecting a suitable Rome-Optimized platform. It is crucial for organizations to partner with vendors that are investing in design and engineering capabilities to produce such a platform at scale.

Equally imperative is the level of trust that firms must place on the vendor’s ability to build and deliver the platform.

Key Capabilities of Trusted Vendors
(top 6 showing)

- Presence of security features: 38
- Ability to physically secure hardware: 34
- Proven track record: 34
- Ability to test, QA and verify: 31
- Ability to procure and verify authenticity: 26
- Investment in R&D: 25

Three Important Factors in Building Trust

Maintaining a secure supply chain. This includes verifying the authenticity of components or parts, procuring components or parts from trusted suppliers, and physically securing the build environment, the system build process, and the process of shipping the system to the customer.

Building security into every design step. This includes incorporating security features in the hardware to prevent malicious attacks as well as developing, testing, and verifying the integrity of the firmware prior to installing it.

Providing proactive updates and system recovery. This includes proactively patching firmware bugs or vulnerabilities and ensuring that updates are delivered and applied in a timely fashion. This also includes providing a restore to a known good state when a system recovery is necessary.

Why wait? Trusted and reputable vendors often take their time to design, build, and deliver a quality product. They focus on getting things right vs. being first to market. It is worth investigating what such vendors have planned for release when investing in new infrastructure.
How to Benefit From Investing in “Right-Fit” Infrastructure

Essential Guidance

• “Right-fit” IT infrastructure provides a solid foundation on which firms can accelerate their digital journey. IT can enable their businesses to produce consistent outcomes by making timely investments in a workload-optimized and innovation-packed infrastructure platform, from a trusted vendor.

• Additionally, IDC studies consistently find that infrastructure operating costs continue to spiral upwards if businesses do not upgrade in a timely manner.

• One-size-fits-all approaches to infrastructure platforms can potentially slow down the pace of business innovation due to performance and scalability. At the same time, an aging infrastructure can quickly become a liability for an organization.

• Older or generic infrastructure platforms are vulnerable to data breaches. Recent disclosures have revealed that vulnerabilities whose root lies in the processor design are difficult to patch without incurring heavy penalties.

• Organizations must therefore discard the “one-size-fits-all” dictum and replace it with a “workload-optimized approach” to making infrastructure decisions – which are based on technology and business economics. They must rely on a trusted vendor as a partner to deliver a fully-optimized and secure platform. Only then can they enable business outcomes at the pace of innovation.

Breaking the “good enough” barrier: Investing in infrastructure is a business decision

Server operating costs in years 4–6 of deployed life are more than 10x higher than the initial acquisition cost of the server.

Compared with continuing to operate installed servers, refreshed servers deliver:

59% lower three-year cost of operations

$4.66M additional revenue per year per organization ($123,400 per server)

Potential benefits of a regular, faster refresh cycle (two three-year life cycles versus one six-year life cycle):

33% lower net cash flow

$14.6M cash flow over six years per 300 servers

<1 YEAR return on investment
**IDC Methodology**

This IDC InfoDoc provides a summary of an extensive validation process performed by IDC in collaboration with the vendor’s teams. IDC relied on the data from the vendor and its own independent research to make statements in this document.

This document is meant to provide a quick set of inferences and insights for IT professionals and business decision makers seeking to perform further due diligence on the capabilities of the product and/or services that have been evaluated in this InfoDoc. However, the goal of this InfoDoc is not to supply detailed hands-on test plans and validation jobs. It is not meant to replace the evaluation process that most businesses will conduct before making any decision to purchase the product and/or services. It is for this reason that this InfoDoc is not designed to be an all-inclusive document on all the capabilities of the product, but rather as a concise document that highlights features/functions of products, their relative performance with respect to a traditional environment and the value these features bring to businesses looking to solving certain problems for Hadoop workloads.

Finally, even though this InfoDoc is a sponsored document, it is not meant to be an IDC endorsement of the product, service or the sponsoring supplier. IDC’s opinions are its own and not influenced by the production of this document.