

AMD EPYC™ 7002 SERIES PROCESSORS AND CLOUDERA®: BIG DATA ANALYTICS FOR ON-PREMISE, CLOUD AND HYBRID DEPLOYMENTS

AUGUST 2019

High Performance

Performance demands are increasing; datacenter space is not. Built on 7nm technology, AMD EPYC™ 7002 Series processors pack more power into the same space as the previous generation. Incredibly high core counts and memory capacity, ample memory bandwidth, and massive I/O combine in the right ratios to help enable breakthrough performance.

Standards Based

Continuing AMD's commitment to industry standards, AMD EPYC™ 7002 Series processors offer you a choice in x86 architecture. x86 compatibility means you can run your x86 based applications on AMD EPYC processors.

Broad Partner Ecosystem

AMD's broad partner ecosystem and collaborative engineering provide tested and validated solutions that help lower your risk and total cost of ownership.

Dedicated Security Processor

AMD EPYC's dedicated security processor manages secure boot, memory encryption, and secure virtualization directly on the SoC. Encryption keys never leave the processor where they can be exposed to intruders.

Cloudera SDX: Shared Data Experience for On-Premise, Cloud and Hybrid Deployments

Cloudera Shared Data Experience (SDX) creates a seamless integration of all analytic disciplines, on-premise, cloud and hybrid.

AMD EPYC is Cloudera Certified Technology

The Cloudera Certified Technology program is designed to make choosing the right technology easier. It means the product has been tested and validated to work with CDH – an enterprise ready distribution of Apache Hadoop and related projects.

Next Gen AMD EPYC™ 7002 Processors: Architectural Innovations Continue to Deliver Exceptional Performance

The second generation of the AMD EPYC™ processor extends AMD's innovation leadership while continuing the philosophy of choice without restriction.

Built with leading-edge 7nm technology, the AMD EPYC™ SoC offers a consistent set of features across a range of choices from 8 to 64 cores, including 128 lanes of PCIe® Gen 4², 8 memory channels and access to up to 4 TB of high-speed memory.

7 nm	PCIe® Gen 4	DDR4 3200
64 Cores per socket	128 PCIe® Gen 4 lanes per socket	8 Memory channels per socket
World's first 7 nm x86 server CPU Highest available core count ¹ to maximize parallelism	World's first PCIe® Gen 4 ready x86 server CPU Doubles the bandwidth of the previous generation	World's first x86 architecture with DDR4 3200 ² Up to 4 TB of memory capacity per socket

AMD EPYC™ 7002 Series processors are offered in single and dual socket configurations.

Single socket enables you to right size the system to your workload needs, such as massive scale-out data management systems, software defined storage, branch office applications, among others.

Dual socket systems offer exceptional performance and density for high-performance computing, virtualization and consolidation, and a variety of private and public cloud applications.

Cloudera Data Hub and Shared Data Experience

Cloudera Enterprise Data Hub provides a complete multi-discipline big data solution. This solution integrates data warehousing, data science, data engineering, search, streaming, and real-time analytics into a unified platform for single source of truth. Cloudera Shared Data Experience (SDX) creates a seamless integration of all analytic disciplines with full security, governance and administration across any set of deployments: on-premise, cloud and hybrid.

AMD EPYC for Big Data Analytics

The advent of big data revolutionized analytics and data science by allowing enterprises to store, access and analyze massive amounts of data of almost any type from any source. This created the need for a new generation of data processing architecture; one where all the data is stored in a vast, ever-expanding data lake, then accessed by various applications using new technologies appropriate to the need.

The EPYC Advantage

Outstanding performance and scalability - Scale-up and scale-out. AMD and its ecosystem partners offer high-performance network connectivity options for massive scale-out systems such as Apache™ Hadoop®.

These needs included traditional historical analysis, now done with the distributed processing methods of big data technologies, as well as integration with existing enterprise systems now able to take advantage of this new source of business insight.

In addition, new capabilities came online in the form of real-time streaming analytics, and ad hoc, interactive analysis on enormous datasets. Click stream analysis, log analytics, recommendation engines are just a few examples of the new generation of applications now in wide use. Some of these applications need massive data injections and high-speed I/O, while others need large amounts of memory - all need the capabilities enabled in the underlying hardware and the software.

The EPYC Advantage

Performance - The AMD EPYC SoC brings new balance to the datacenter. The highest core count yet in an AMD x86-architecture server processor, large memory capacity, memory bandwidth and I/O density are all brought together with the right ratios to help performance reach new heights.

AMD has partnered with leading software and server vendors to create certified reference architectures for customers to demonstrate strong performance, lower risk and to help reduce implementation cost.

Cloudera Enterprise

Cloudera is a thought leader in the creation and application of enterprise-wide big data analytics solutions enabling organizations to answer their most valuable questions. Cloudera Enterprise Data Hub brings together data warehousing, data science, data engineering, search, streaming and real-time analytics into a unified platform with a single source of truth.

Organizations now run diverse, multidisciplinary big data workloads that span analytic databases, operational databases, data engineering applications, and data science applications. Many of these workloads operate on the same underlying data.

Recognizing that most business applications require a combination of core functions to solve real-world problems, Cloudera identified then integrated these functions enabling them to operate on the same data set.

- **Data Engineering** enables the business to run batch or stream processes that speed ETL and train machine learning models.
- **Analytic Database** delivers fast time-to-insight with the flexibility and agility to run in any environment and against any type of data.
- **Data Science** enables the business to do exploratory data science at big data scale with full data security and governance.
- **Operational Database** enables the business to build data-driven applications that deliver near real-time insights.

This integration solves a thorny problem, but it exists within the context of a larger issue. Compute has moved to the data and is stateless, either transient or long running, but the data may be on-premise or in the cloud, or both.

Further, the meta-data that describes the data: table definitions, governance artifacts, metadata classifications, access permissions, and business definitions, often called the ‘data context,’ is difficult to keep consistent across the increasing number of workloads and data locations.

Cloudera SDX: Shared Data Experience

Cloudera SDX offers a modular software framework that ensures a shared data experience across all deployment types, including multiple public clouds, private cloud, hybrid cloud, and bare metal configurations. SDX applies stateful, centralized, and consistent data context services making it possible for hundreds of different workloads to run against shared or overlapping sets of data. SDX makes multi-disciplinary data applications easier to develop, less expensive to deploy and increasingly important in today’s environment, more consistently secure.

SDX is comprised of five discrete functions that together solve a difficult problem – providing a shared data experience for a platform that supports a diverse set of workloads and user interaction models. See Figure 1.

- **Shared Governance** provides the ability to govern the data in a unified manner so users can easily discover new data, understand where that data came from, and track how it has been modified.
- **Shared Workload Management** enables administrators to create, manage and optimize workloads individually or as a collection and to allocate resources and assign workload priority based on business requirements.
- **Shared Ingest & Replication** provide the ability to ingest data once and make it available to all applications and users without additional ingest pipelines or copies of data, and to replicate data on demand to remote locations or directly to the cloud.
- **Shared Data Catalog** provides a common catalog of schema and lineage metadata to each workload and user accessing the platform for maximum efficiency and productivity.
- **Shared Security** implements consistent, granular authentication, authorization, encryption, and compliance controls in a unified manner across the entire platform.

The EPYC Advantage

Security – AMD developed a suite of processor-based encryption and security features designed to guard the processor from internal and external attacks. This data encryption limits risks and provides protection from data breaches.

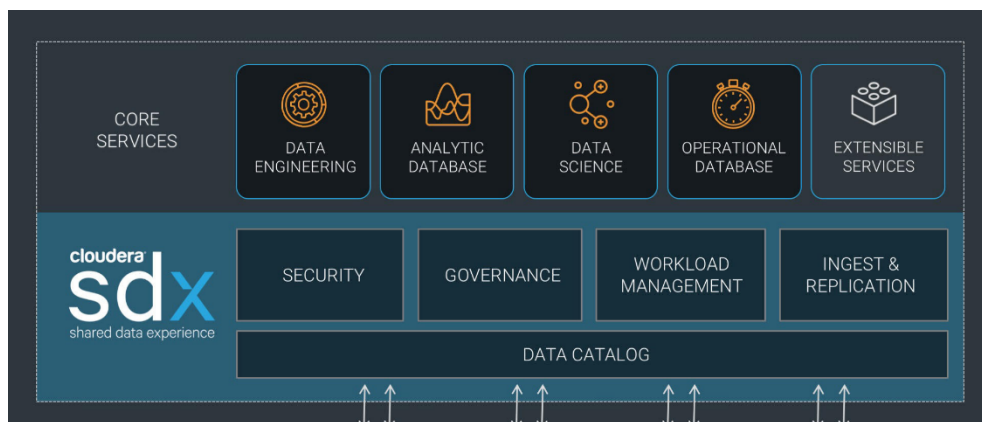


Figure 1: Cloudera SDX

Reference Architecture for AMD EPYC and Cloudera Enterprise

The reference architectures for AMD EPYC processors and Cloudera Enterprise provide options for the performance and scalability requirements needed to maximize the investment in Big Data Analytics.

AMD EPYC's ability to provide a no compromise single-socket solution ensures you are only paying for the processing power the application needs. A single-socket server has all the I/O and memory bandwidth available in a dual-socket server without the extra cost. For high-performance applications, AMD EPYC-based dual-socket servers offer phenomenal core density.

These reference architectures offer the foundation for creating custom configurations that meet unique application demands with the various storage and connectivity options to choose from our leading server OEM partners.

NO COMPROMISE SINGLE-SOCKET		VERSATILE DUAL-SOCKET	
Management Nodes: 3		Management Nodes: 3	
CPU	1 x AMD EPYC 7502P, 7552, or 7702P*	CPU	2 x AMD EPYC 7402, 7452, or 7552*
Cores	32/48/64	Cores	24/32/48 per CPU, 48/64/96 per node
Threads	64/96/128	Threads	48/64/96 per CPU, 96/128/192 per node
Memory	256 GB	Memory	256 GB
Network	10/25/40/100 Gigabit Ethernet	Network	10/25/40/100 Gigabit Ethernet
Storage: OS	2 x SSDs or 2 x HDDs	Storage: OS	2 x SSDs or 2 x HDDs
Storage: Data	2 x NVMe drives, or 4 x SAS/SATA SSDs, or 8 x SAS/SATA HDDs	Storage: Data	2 x NVMe drives, or 4 x SAS/SATA SSDs, or 8 x SAS/SATA HDDs
Data Nodes: Minimum 3, scalable to hundreds		Data Nodes: Minimum 3, scalable to hundreds	
CPU	1 x AMD EPYC 7502P, 7552, or 7702P*	CPU	2 x AMD EPYC 7402, 7452, or 7552*
Cores	32/48/64	Cores	24/32/48 per CPU, 48/64/96 per node
Threads	64/96/128	Threads	48/64/96 per CPU, 96/128/192 per node
Memory	128/256 GB (supports up to 2 TB)	Memory	128/256 GB (supports up to 4 TB)
Network	10/25/40/100 Gigabit Ethernet	Network	10/25/40/100 Gigabit Ethernet
Storage: OS	2 x SSDs or 2 x HDDs	Storage: OS	2 x SSDs or 2 x HDDs
Storage: Data		Storage: Data	
High Performance	4-8 x NVMe drives or 8-12 x SAS/SATA SSDs	High Performance	4-8 x NVMe drives or 8-12 x SAS/SATA SSDs
Performance	24 x SAS/SATA HDDs (SFF)	Performance	24 x SAS/SATA HDDs (SFF)
Capacity	12-16 x SAS/SATA HDDs (LFF)	Capacity	12-16 x SAS/SATA HDDs (LFF)

*other AMD EPYC 7002 series processors are also supported

Conclusion

Versatility and agility are among the most important requirements in modern datacenters. The AMD EPYC system-on-a-chip (SoC) enables organizations to deploy systems that precisely meet today's needs while positioning themselves for tomorrow's requirements.

Cloudera delivers on the promise of complete integration of different analytic disciplines for big data: operations, data science, data engineering and machine learning, all optimized for on-premise, cloud and hybrid deployments.

Together, AMD and Cloudera empower the development of modern data applications that implement solutions to a diverse set of business problems with high-performance processing and cost-effective solutions that are perfectly sized for current needs and easily scalable as your business grows.

For more information about AMD's EPYC line of processors visit: <http://www.amd.com/epyc>

For more information about Cloudera visit: <http://www.cloudera.com>.

For more information about Cloudera Certified Technology program visit:
<https://www.cloudera.com/partners/partners-listing.html>

FOOTNOTES

1. Best-in-class based on industry-standard pin-based (LGA) X86 processors. NAP-166.
2. Some supported features and functionality of second-generation AMD EPYC™ processors (codenamed "Rome") require a BIOS update from your server manufacturer when used with a motherboard designed for the first-generation AMD EPYC 7000 series processor. A motherboard designed for "Rome" processors is required to enable all available functionality. ROM-06.

©2019 Advanced Micro Devices, Inc. All rights reserved. AMD, the AMD Arrow logo, EPYC and combinations thereof are trademarks of Advanced Micro Devices, Inc. Cloudera and the Cloudera logo are registered trademarks or registered trademarks of Cloudera, Inc. in the United States and other countries. PCIe and PCI Express are registered trademarks of PCI-SIG Corporation. Apache and Apache Hadoop are either registered trademarks or trademarks of the Apache Software Foundation in the United States and other countries. Other product names used in this publication are for identification purposes only and may be trademarks of their respective companies.

The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions and typographical errors. AMD reserves the right to revise this information and to make changes from time to time to the content hereof without obligation of AMD to notify any person of such revisions or changes.