AMD Hardware and Software Platforms for HPC and Machine Intelligence

Dr. Timour Paltashev
Senior Manager
Radeon Technologies Group

Parallel Computational Technologies PCT’2018, Rostov-on-Don, Russia
OUTLINES

1. New server processor EPYC for enterprises and data centers
2. New architecture features of EPYC built on Zen CPU core
3. EPYC memory hierarchy and virtualization protection
4. Capable I/O and power management subsystems
5. IHV and ISV platforms ecosystem for EPYC
6. Radeon Instinct GPU accelerators combination with EPYC
7. P47 systems with petaflop range performance
8. Radeon Open Compute (ROCm) open source SW stack for HPC and MI
9. Deep learning applications
10. Summary
AMD EPYC™ 7000 SERIES PROCESSOR

Architected for Enterprise and Datacenter Server TCO
Up to 32 High-Performance “Zen” Cores
8 DDR4 Channels per CPU
Up to 2TB Memory per CPU
128 PCIe® Lanes
Dedicated Security Subsystem
Integrated Chipset for 1-Chip Solution
Socket-Compatible with Next Gen EPYC Processors
### WHAT EPYC OFFERS

- **Industry leading core count**
  - Up to 32 cores per socket for high parallelism
- **Largest memory capacity**
  - Up to 2TB per socket for in-memory processing
- **High memory throughput**
  - Over 290MB/s as measured by STREAM Triad
- **Industry leading I/O bandwidth**
  - 128 PCIe Gen 3 lanes allow for fast network connectivity
- **Embedded security**

### WHAT EPYC DELIVERS

- **High parallelism to quickly crunch through large datasets**
- **Blazing fast access to large datasets**
- **Capacity to load large datasets in-memory for faster processing**
- **Faster load/store of persistent data**
- **Data Security**
### Key Tenets of EPYC™ Processor Design

<table>
<thead>
<tr>
<th>Category</th>
<th>Tenets</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CPU Performance</strong></td>
<td>“Zen”—high performance x86 core with Server features</td>
</tr>
<tr>
<td><strong>Virtualization, RAS, and Security</strong></td>
<td>ISA, Scale-out microarchitecture, RAS</td>
</tr>
<tr>
<td></td>
<td>Memory Encryption, no application mods: security you can use</td>
</tr>
<tr>
<td><strong>Per-socket Capability</strong></td>
<td>Maximize customer value of platform investment</td>
</tr>
<tr>
<td></td>
<td>MCM – allows silicon area &gt; reticle area; enables feature set</td>
</tr>
<tr>
<td></td>
<td>MCM – full product stack w/single design, smaller die, better yield</td>
</tr>
<tr>
<td><strong>Fabric Interconnect</strong></td>
<td>Cache coherent, high bandwidth, low latency</td>
</tr>
<tr>
<td></td>
<td>Balanced bandwidth within socket, socket-to-socket</td>
</tr>
<tr>
<td><strong>Memory</strong></td>
<td>High memory bandwidth for throughput</td>
</tr>
<tr>
<td></td>
<td>16 DIMM slots/socket for capacity</td>
</tr>
<tr>
<td><strong>I/O</strong></td>
<td>Disruptive 1 Processor (1P), 2P, Accelerator + Storage capability</td>
</tr>
<tr>
<td></td>
<td>Integrated Server Controller Hub (SCH)</td>
</tr>
<tr>
<td><strong>Power</strong></td>
<td>Performance or Power Determinism, Configurable TDP</td>
</tr>
<tr>
<td></td>
<td>Precision Boost, throughput workload power management</td>
</tr>
</tbody>
</table>
## VIRTUALIZATION ISA AND MICROARCHITECTURE ENHANCEMENT

<table>
<thead>
<tr>
<th>ISA FEATURE</th>
<th>NOTES</th>
<th>“Bulldozer”</th>
<th>“Zen”</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Poisoning</td>
<td>Better handling of memory errors</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>AVIC</td>
<td>Advanced Virtual Interrupt Controller</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Nested Virtualization</td>
<td>Nested VMLOAD, VMSAVE, STGI, CLGI</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Secure Memory Encryption (SME)</td>
<td>Encrypted memory support</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Secure Encrypted Virtualization</td>
<td>Encrypted guest support</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>(SEV)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PTE Coalescing</td>
<td>Combines 4K page tables into 32K page size</td>
<td>✓</td>
<td></td>
</tr>
</tbody>
</table>

### Microarchitecture Features

<table>
<thead>
<tr>
<th>Reduction in World Switch latency</th>
<th>Dual Translation Table engines for TLBs</th>
<th>Tightly coupled L2/L3 cache for scale-out Private L2: 12 cycles Shared L3: 35 cycles</th>
<th>Cache + Memory topology reporting for Hypervisor/OS scheduling</th>
</tr>
</thead>
</table>

“Zen” Core built for the Data Center

Parallel Computational Technologies PCT’2018, Rostov-on-Don, Russia
HARDWARE MEMORY ENCRYPTION
SECURE ENCRYPTED VIRTUALIZATION (SEV)

- Protects against physical memory attacks
- Single key is used for encryption of system memory
  - Can be used on systems with Virtual Machines (VMs) or Containers
- OS/Hypervisor chooses pages to encrypt via page tables or enabled transparently
- Support for hardware devices (network, storage, graphics cards) to access encrypted pages seamlessly through DMA
- No application modifications required

Defense Against Unauthorized Access to Memory
HARDWARE MEMORY ENCRYPTION
SECURE ENCRYPTED VIRTUALIZATION (SEV)

- Protects VMs/Containers from each other, administrator tampering, and untrusted Hypervisor
- One key for Hypervisor and one key per VM, groups of VMs, or VM/Sandbox with multiple containers
- Cryptographically isolates the hypervisor from the guest VMs
- Integrates with existing AMD-V technology
- System can also run unsecure VMs
- No application modifications required
BUILT FOR SERVER: RAS (RELIABILITY, AVAILABILITY, SERVICEABILITY)

- **Caches**
  - L1 data cache SEC-DED ECC
  - L2+L3 caches with DEC-TED ECC

- **DRAM**
  - DRAM ECC with x4 DRAM device failure correction
  - DRAM Address/Command parity, write CRC—with replay
  - Patrol and demand scrubbing
  - DDR4 post package repair

- **Data Poisoning and Machine Check Recovery**

- **Platform First Error Handling**

- **Infinity Fabric link packet CRC with retry**

Enterprise RAS Feature Set
I/O SUBSYSTEM

- 8x16 links available in 1 and 2 socket systems
  - Link bifurcation support; max of 8 PCIe® devices per x16
  - Socket-to-Socket Infinity Fabric, PCIe, and SATA support
- 32GB/s bi-dir bandwidth per link, 256GB/s per socket
- PCIe Features*
  - Advanced Error Reporting (AER)
  - Enhanced Downstream Port Containment (eDPC)
  - Non-Transparent Bridging (NTB)
  - Separate RefClk with Independent Spread support (SRIS)
  - NVMe and Hot Plug support
  - Peer-to-peer support
- Integrated SATA support

Architected for Massive I/O Systems with Leading-edge Feature Set, Platform Flexibility
PER CORE LINEAR VOLTAGE REGULATION

- Running all dies/cores at the voltage required by the slowest wastes power
- Per-core voltage regulator capabilities enable each core’s voltage to be optimized
  - Significant core power savings and variation reduction
LEADERSHIP DUAL-SOCKET

WORLD RECORD BENCHMARKS!

EPYC 7601

SPECfp_rate2006
SPECfp_rate2017

EPYC 7601

14% More Cores

33% More Memory Bandwidth

2.6x More Memory Capacity

UP TO 2.6x More Performance / $

Feature and perf/$ comparison to 2 Intel Xeon Platinum 8180. Perf/$ based on published prices and published SPECfp_rate2016 scores on spec.org. See Endnotes.

Parallel Computational Technologies PCT’2018, Rostov-on-Don, Russia
FLOATING POINT LEADERSHIP IN 3RD PARTY TESTING

VS XEON PLATINUM 8176

Note: All benchmarks taken from Anandtech.com

COMPELLING FLOATING POINT PERFORMANCE

- More high performance cores
- Dedicated and efficient FP engine
- More Memory Bandwidth

Parallel Computational Technologies PCT'2018, Rostov-on-Don, Russia
THE ECOSYSTEM IS READY

<table>
<thead>
<tr>
<th>OEM / ODM</th>
<th>CLOUD</th>
<th>INTEGRATORS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hewlett Packard Enterprise</td>
<td>Microsoft Azure</td>
<td>AMAX</td>
</tr>
<tr>
<td>Sugon</td>
<td>GIGABYTE</td>
<td>PENGUIN COMPUTING</td>
</tr>
<tr>
<td>Lenovo</td>
<td>ASUS</td>
<td>E4</td>
</tr>
<tr>
<td>TYAN</td>
<td>Wlstron</td>
<td>EchoStreams</td>
</tr>
<tr>
<td>Inventec</td>
<td>HBC</td>
<td>Clustervision</td>
</tr>
<tr>
<td>SUPERMICRO</td>
<td>Quanta Computer</td>
<td>Siemens</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
THE EPYC Ecosystem
BUILDING ON THE UBIQUITY AND ROBUSTNESS OF X86

<table>
<thead>
<tr>
<th>Hypervisor</th>
<th>Operating Systems</th>
<th>Tools</th>
</tr>
</thead>
<tbody>
<tr>
<td>Microsoft Hyper-V</td>
<td>redhat</td>
<td>KVM</td>
</tr>
<tr>
<td>GCC</td>
<td>php</td>
<td>SUSE</td>
</tr>
<tr>
<td>docker</td>
<td>vmware</td>
<td>CITRIX</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>IHV Partners</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mellanox</td>
</tr>
<tr>
<td>Micron</td>
</tr>
<tr>
<td>NVIDIA</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Infrastructure</th>
<th>Storage</th>
</tr>
</thead>
<tbody>
<tr>
<td>SharePoint</td>
<td>Exchange</td>
</tr>
<tr>
<td>Apache</td>
<td>ceph</td>
</tr>
</tbody>
</table>

Parallel Computational Technologies PCT'2018, Rostov-on-Don, Russia
ARCHITECTED TOGETHER
TO DELIVER BREAKTHROUGH PERFORMANCE

AI Cloud | Machine Intelligence | Deep Learning | High Performance Computing

Parallel Computational Technologies PCT’2018, Rostov-on-Don, Russia
Radeon Instinct™ MI25

- 64 Next-Gen Compute Units (NGC – 4096 Stream Processors)
- 12.3 TFLOPS Peak Single Precision Compute (FP32)
- 24.6 TFLOPS Peak Half Precision Compute (FP16)
- 16 GB HBM2 High Bandwidth Cache
- 484 GB/s Memory Bandwidth

Parallel Computational Technologies PCT’2018, Rostov-on-Don, Russia
Radeon Instinct On EPYC Platform

- Direct Device Connect
- Lowest System Cost
- Lower Latency Architecture
- Peer to Peer Communication
- High Density Footprint

Optimized for VDI, ML & HPC Computing

Parallel Computational Technologies PCT’2018, Rostov-on-Don, Russia
AMD EPYC™ AND RADEON INSTINCT™

SOLUTIONS FOR DEEP LEARNING

EPYC provides:
- Up to 128 high bandwidth PCIe® gen 3.0 lanes for unmatched connectivity
- 8 memory channels for high bandwidth and capacity
- 32 high performance cores

Radeon Instinct MI25 delivers:
- Hyperscale and HPC-class heterogeneous compute for Machine Intelligence and Deep Learning, along with HPC workloads
- 4,096 Stream Processors for 24.6 TFLOPS (FP16); 12.3 TFLOPS (FP32); 768 GFLOPS (FP64)
- 16GB of latest HBM2 ECC GPU memory with 484 GB/s of memory bandwidth*
- Radeon Instinct’s open ecosystem approach to datacenter design with ROCm open software platform and MIOpen libraries

Combining these technologies, they deliver a world class machine learning platform...

*ECC support is limited to the HBM2 memory and ECC protection is not provided for internal GPU structures.
HARDWARE PLATFORM

Inventec P47

- AMD EPYC™ 7601 Processor
  - 32 cores (64 threads)
  - 2200Mhz base / 3200Mhz boost
  - 512 GB physical memory

4 x Radeon Instinct™ MI25 Accelerators

- “Vega10” GPU Arch
- 16 GB HBM2
- 12.3 TFLOPS peak single precision
RCCL – RADEON COMMUNICATION COLLECTIVE LIBRARY

Designed as Library of Communication Primitives for functions like all-gather, reduce, broadcast for multi-GPU high performance inter-communication

Inspired by MPI Communication Collective Libraries and NCCL

Configuration supported
- all-reduce
- all-gather
- reduce-scatter
- reduce
- Broadcast

- Fully Open Source
ROCM FOR DISTRIBUTED SYSTEMS

ROCM Has Built In Support for Distributed Computing:

CPU directly accesses GPU memory

- Expose entire GPU frame buffer as addressable memory through PCIe BAR (Large Bar)
- Map GPU pages to CPU pages
- Allow CPU to directly load/store from/to GPU memory

ROCNRDMA Peer to Peer HCA directly access GPU memory:

- Supports Mellanox PeerDirect
- Allow IB HCA to directly read/write data from/to GPU memory

Available and when installing OFED on ROCm, Peer Sync enabled by default in ROCm

Parallel Computational Technologies PCT'2018, Rostov-on-Don, Russia
Project 47  1 PETAFLOP
(GPGPU RACK)
SINGLE PRECISION

1,028 Single Precision TFLOPS
2,056 Half Precision TFLOPS

30.05
GIGAFLOPS/WATT
(at single precision)

20 x AMD EPYC 7601 CPU
80 x Radeon MI-25 Instinct Cards
10.24 Terabytes of Samsung DDR4 2667 Memory
20 x Mellanox 100G IB cards + 1 switch

Inventec P47 & Project 47 Racks Distributed by AMAX

Parallel Computational Technologies PCT’2018, Rostov-on-Don, Russia
ROCM 1.7.x
Open Source Software Platform GPU Computing

- Radeon Instinct Enablement
- Multi-GPU Connectivity
- Improved System Management
- Richer Programming Support

Parallel Computational Technologies PCT’2018, Rostov-on-Don, Russia
 ROCm PROGRAMMING MODEL OPTIONS

**HIP**

*Convert CUDA to portable C++*

- Single-source Host+Kernel
- C++ Kernel Language
- C Runtime (CUDA-like)
- Platforms: AMD GPU, NVIDIA (same perf as native CUDA)

When to use it?
- Port existing CUDA code
- Developers familiar with CUDA
- New project that needs portability to AMD and NVIDIA

**HCC**

*True single-source C++ accelerator language*

- Single-source Host+Kernel
- C++ Kernel Language
- C++ Runtime
- Platforms: AMD GPU

When to use it?
- New projects where true C++ language preferred
- Use features from latest ISO C++ standards

**OpenCL**

*Khronos Industry Standard accelerator language*

- Split Host/Kernel
- C99-based Kernel Language
- C Runtime
- Platforms: CPU, GPU, FPGA

When to use it?
- Port existing OpenCL code
- New project that needs portability to CPU, GPU, FPGA
# SOFTWARE STACK

<table>
<thead>
<tr>
<th>Applications</th>
<th>Machine Learning Apps (ResNet50 Training, CANDLE, etc)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frameworks</td>
<td>Caffe</td>
</tr>
<tr>
<td>Middleware &amp; Libraries</td>
<td>MIOpen</td>
</tr>
<tr>
<td>Programming Models</td>
<td>HCC</td>
</tr>
<tr>
<td>ROCm</td>
<td>ROCm Platform</td>
</tr>
</tbody>
</table>

*Italics = Under Dev by AMD*

Parallel Computational Technologies PCT'2018, Rostov-on-Don, Russia
# OPEN AND SHUT: THE CASE FOR AMD’S OPEN SOURCE MACHINE INTELLIGENCE SOFTWARE STACK

<table>
<thead>
<tr>
<th>Open-source MI Software</th>
<th>AMD / ROCm</th>
<th>NVIDIA / CUDA</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAFFE</td>
<td>Open-source</td>
<td>Open-source</td>
</tr>
<tr>
<td>TensorFlow</td>
<td>Open-source</td>
<td>Open-source</td>
</tr>
<tr>
<td>Programming Language</td>
<td>Open-source (HIP)</td>
<td>Proprietary (CUDA)</td>
</tr>
<tr>
<td>Accelerated MI Library</td>
<td>Open-source (MIOpen)</td>
<td>Proprietary (cuDNN)</td>
</tr>
<tr>
<td>Accelerated Math Libs</td>
<td>Open-Source (rocBLAS, rocRand, rocFFT, sparse)</td>
<td>Proprietary (cuBLAS, cuRNG, cuFFT, cuSparse)</td>
</tr>
<tr>
<td>Communication Library</td>
<td>Open-source (RCCL)</td>
<td>NCCL</td>
</tr>
<tr>
<td>Runtime</td>
<td>Open-source (ROCr)</td>
<td>Proprietary</td>
</tr>
<tr>
<td>Linux Driver</td>
<td>Open-source (AMDGPU)</td>
<td>Proprietary</td>
</tr>
<tr>
<td>Documented ISA</td>
<td>Open (GCN)</td>
<td>Proprietary</td>
</tr>
</tbody>
</table>

Italics = Under Dev

Software: [https://rocm.github.io](https://rocm.github.io)

Parallel Computational Technologies PCT’2018, Rostov-on-Don, Russia
Deep Learning

Vision
- Convolution Neural Networks (CNNs)
  - Convolution
  - 4D/5D Tensors

First Focus Area

Natural Language Processing
- Recurrent Neural Networks (RNNs)
  - Long Short Term Memory (LSTM)
  - Recurrent Weighted Average (RWA)
  - Gated Recurrent Unit (GRU)

Creative / Higher Reasoning
- Generative Adversarial Networks (GANs)
  - Reinforcement Learning

Development Plan

Recognizing cats from YouTube and road scenes

“Girl in pink dress is jumping in air”

Beating the world champion in Go

Parallel Computational Technologies PCT’2018, Rostov-on-Don, Russia
SUMMARY

• EPYC processor is available in multiple configurations (8-32 cores) and dual sockets (up to 64 cores) in the systems from major industrial HPC vendors

• Probably best customized solution for cloud and enterprise servers

• Excellent SW ecosystem with hypervisors, OS, tools, databases, analytics, infrastructure and storage support

• AMD ROCm and Radeon Instinct™ platform available now
  — Provides new option for machine intelligence TensorFlow and Keras Applications

• Delivers competitive performance and excellent scaling

• Continuously developing ROCm open source software ecosystem
  — Path to more performance, features, libraries, and applications
BACKUP
Keras is a high-level neural net API

- Designed for rapid prototyping and experimentation
- Create layers with single line of code
- Runs on top of TensorFlow (also CNTK, Theano)
- Very popular in academic and commercial installations

Keras runs on ROCm on TensorFlow

- Large ecosystem of Keras apps have new platform choice

CANDLE

- DOE Exascale Computing Project
- Uses deep learning to facilitate cancer research

Some models (P1B3) are now running on Keras on ROCm.

```python
import keras
...
x_train, y_train = loadInput()
...
model = Sequential()
model.add(Dense(units=64, input_dim=100))
model.add(Activation('relu'))
model.add(Dense(units=10))
model.add(Activation('softmax'))
```
ROCM SOURCE CODE

Source-Code Repositories for Kernel Driver + Thunk + Run Time
• ROCk: https://github.com/RadeonOpenCompute/ROCK-Kernel-Driver
• ROCt: https://github.com/RadeonOpenCompute/ROCT-Thunk-Interface
• ROCr: https://github.com/RadeonOpenCompute/ROCR-Runtime

HCC Compiler
• HCC compiler source: https://github.com/RadeonOpenCompute/hcc
• HCC Clang source: https://github.com/RadeonOpenCompute/hcc-clang

HIP
• HIP run time and tools for porting Cuda applications with HCC: https://github.com/GPUOpen-ProfessionalCompute-Tools/HIP
• HIP examples: https://github.com/GPUOpen-ProfessionalCompute-Tools/HIP-Examples

LLVM Source for GCN ISA Compiler
• LLVM source code for native GCN ISA code generation: https://github.com/RadeonOpenCompute/llvm
• ROCm device-library compiler intrinsics with Open Compute Math Library and Open Compute kernel language: https://github.com/RadeonOpenCompute/ROCm-Device-Libs

Parallel Computational Technologies PCT’2018, Rostov-on-Don, Russia
DOCUMENTS

- ROCm.github.io
  - https://rocm.github.io/documentation.html
  - HSA Foundation Runtime Spec used by ROCr
  - ROCm error codes https://rocm.github.io/ROCmRTec.html
  - ROCm SMI https://github.com/RadeonOpenCompute/ROC-smi/blob/master/README.md
  - USE of PCIe Atomics and Large Bar Support
    https://github.com/RadeonOpenCompute/RadeonOpenCompute.github.io/blob/master/ROCmPCleFeatures.md
  - ROCmRDMA https://github.com/RadeonOpenCompute/ROCnRDMA
ROCM DOCUMENTATION: COMPILER

- Compiler
  - HCC Overview [https://github.com/RadeonOpenCompute/hcc/wiki](https://github.com/RadeonOpenCompute/hcc/wiki)
  - LLVM Documentation [http://llvm.org/docs/AMDGPUUsage.html](http://llvm.org/docs/AMDGPUUsage.html)