
# Flash Solutions for the Skyrocketing AI/ML Applications Market

Dr. Radoslav Danilak

CEO Tachyum Inc.

Flash Memory Summit

# IDC: AI / ML CAGR 38% - 44%

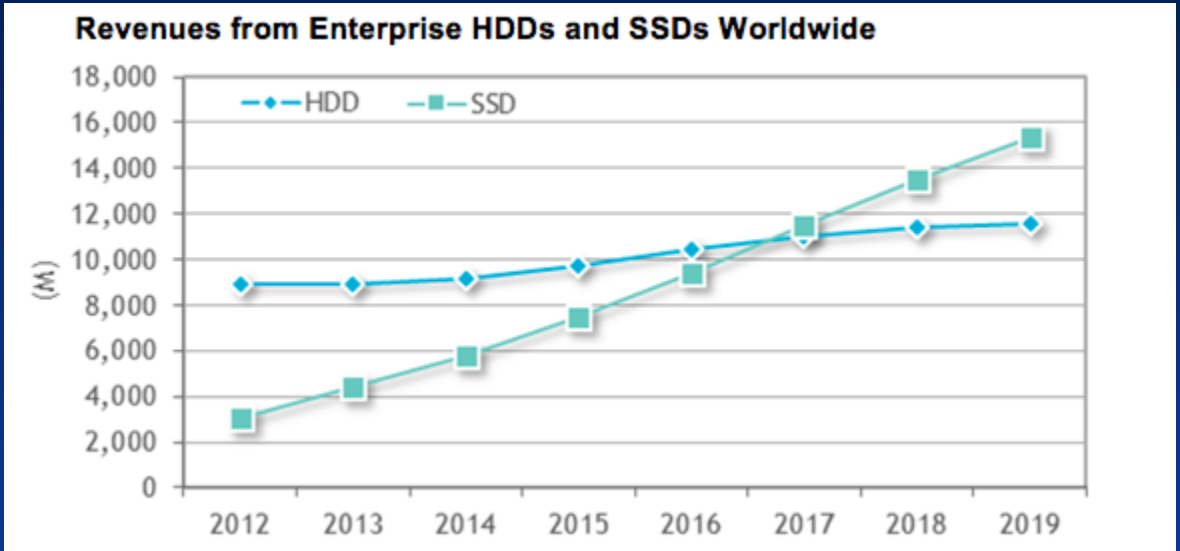
- Worldwide spending on artificial intelligence (AI) systems is forecast to reach $35.8 billion in 2019
  - 44.0% CAGR vs. 2018
- Spending on AI - $79.2 billion in 2022
  - 38% CAGR from 2018-2022
- 2019 top AI Use Cases
  - Automated customer service ($4.5 billion worldwide)
  - Sales process automation ($2.7 billion)
  - Automated threat intelligence and prevention ($2.7 billion)

# AI/ML Driving Flash Memory CAGR

The raw material for AI/ML is data – in vast quantities

Gartner predicts business value created by AI:
- $3.9T in 2022

McKinsey:
82% of enterprises show avg. 17% ROI from their AI/ML investments




Deloitte:
% of companies using AI/ML

- USA/NA: 23%
- EU: 21%
- China: 19%

IDC: all-flash array CAGR - 21.4% (through 2020)

Flash Memory Summit

# Cloud Computing & AI/ML

- AI/ML Hyperscalers need:
  - Low latency access to massive data sets
    - Industry is responding
  - High throughput / low power processing
    - Stay tuned…

# Flash Industry Response

- Intel and Micron's 3D Xpoint
- Samsung Z-NAND 12-20µs latency for random reads and 16µs for random writes
- Toshiba XL-FLASH Low Latency 3D NAND

# NVIDIA Dominates AI/ML h/w

- NVIDIA Q1 2019:
  - 71% growth in its datacenter business
  - Revenue $701M for the quarter
- NVIDIA Pegasus:
  - 2019 Bosch/Daimler ADAS tests, in SJ

**Share of Compute Instance Types with Dedicated Accelerators Offered by the Top Four Public Clouds**
(Alibaba Cloud, Amazon Web Services, Google Cloud & Microsoft Azure)

| Company | Accelerator | March 2019 | April 2019 | May 2019 |
|---|---|---|---|---|
| **NVIDIA** | GPU | 97.0% | 97.3% | 97.4% |
| **AMD** | GPU | 1.2% | 1.1% | 1.0% |
| **Xilinx** | FPGA | 1.1% | 1.0% | 1.0% |
| **Intel** | FPGA | 0.6% | 0.6% | 0.6% |
| **Total Types** | All | 1,852 | 1,990 | 2,003 |

*Source: Liftr Cloud Insights, June 2019*

320 TTFlops ~500 Watts

NVIDIA Pegasus AI Computer

# AI Bottlenecks

- Networking & Storage Stack
  - NMVe has exposed the network and the storage I/O stack as bottlenecks.
  - Vendors now running NVMe over Ethernet, Infiniband, using RDMA
- High speed / low power compute

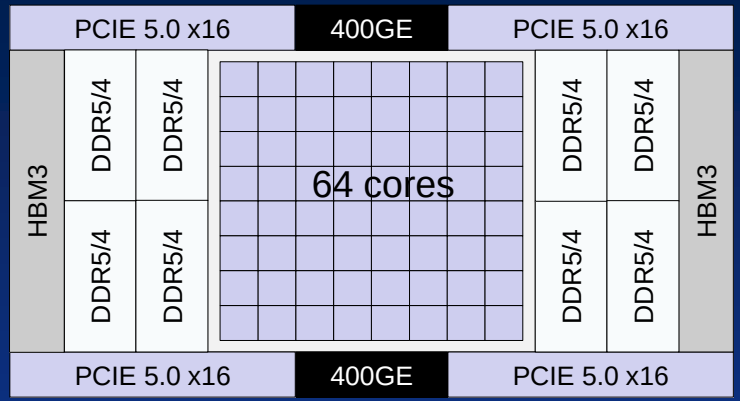
Flash Memory Summit

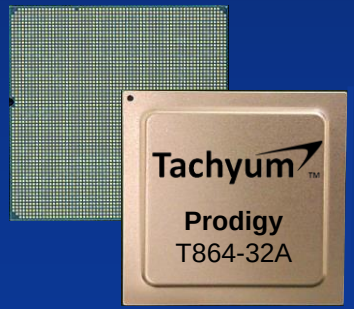
# Prodigy: 1st Universal Processor

- 7nm TSMC FinFET
- 64 cores / >4GHz
- 128 Tensor TeraFlops/Socket @ <200 watts
- Seamlessly switch from DC to AI/ML/HPC workloads

# Prodigy Universal Processor

- 64 cores, each core faster than Xeon
  - 8 DDR5/4
  - 72 PCI Express 5.0
  - 2 x 400/100/50/25/10G Ethernet
  - 2 HBM3 (optional)
  - 32MB fully coherent L3 cache
  - 180W, 64 4GHz cores at 0.825V running AI
- Faster than Xeon, smaller than ARM
  - Data travels over very short wires mitigating the “slow wires” problem
  - Out-of-Order execution with Compiler

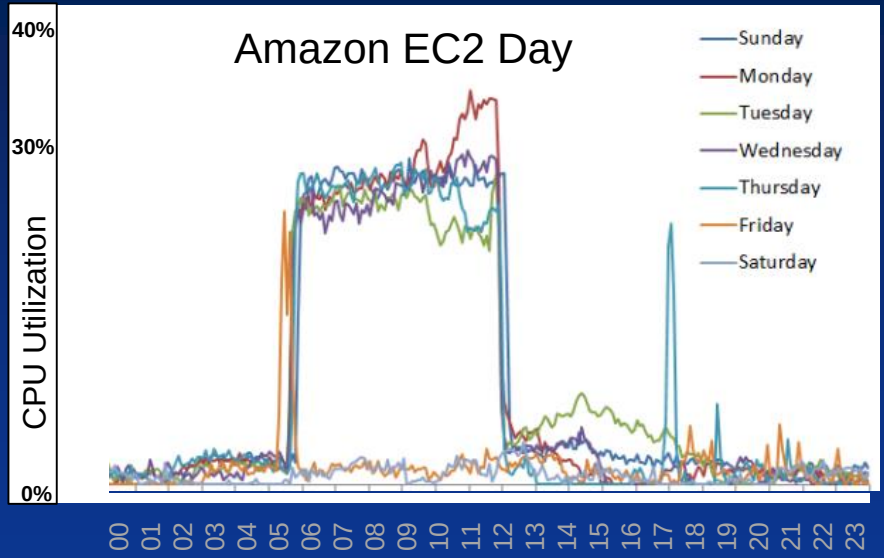


# Big AI for Datacenters – CAPEX Free

Flash Memory Summit

- Existing Processors - too slow for AI therefore, GPU or TPUs are used
- Universal Processor / AI chip:
  - 10x more AI using idle servers
- Prodigy enables idle servers to be seamlessly and dynamically reconfigured into HPC/AI systems
- Prodigy delivers 10x more Data Center AI
  - CAPEX FREE!




Prodigy: Silicon in 2020 - will drive significantly increased SSD sales