DGX SYSTEMS: DEEP LEARNING FROM DESK TO DATA CENTER

Markus Weber and Haiduong Vo
NVIDIA DGX SYSTEMS

Agenda

• NVIDIA DGX-1
• NVIDIA DGX STATION
ONE YEAR LATER - NVIDIA DGX-1
Barriers Toppled, the Unsolvable Solved - a Sampling of DGX-1 Impact

DGX-1 launch | UC Berkley | CSIRO | MIT | CMU | Fidelity | Skymind | RIKEN
April 2016

OpenAI | NYU | Mass. General Hosp. | DFK | IDSIA | Microsoft | SAP | NVIDIA SATURNV launch | Facebook

October 2017
INTRODUCING THE DGX FAMILY

AI WORKSTATION

DGX Station

with

Tesla V100
The Personal AI Supercomputer

AI DATA CENTER

DGX-1

with

Tesla P100
The World’s First AI Supercomputer in a Box

with

Tesla V100
The Essential Instrument for AI Research

CLOUD-SCALE AI

NVIDIA GPU Cloud

Cloud platform with the highest deep learning efficiency
NVIDIA DGX-1 OVERVIEW
## USERS AND CUSTOMERS FOR NVIDIA DGX-1

### DATA SCIENTISTS AND AI RESEARCHERS

**Why use NVIDIA DGX-1?**
- Reduce DL training time
- Analyze and visualize vast amount of data
- Accelerate deep learning frameworks
- Design more sophisticated neural networks

### CIO, CTO, CMO, LINE OF BUSINESS (LOB)

**Why buy NVIDIA DGX-1?**
- Extract actionable insights
- Create new business opportunities
- Turn huge amounts of data into extreme value

### IT DIRECTORS AND MANAGERS

**Why add NVIDIA DGX-1 into your datacenter?**
- Cut infrastructure footprint by 400x and reduce cost by 20x
- Reduce power and cooling costs
- Save installation and configuration time
NVIDIA DGX-1 WITH VOLTA
Highest Performance, Fully Integrated HW System

1 PetaFLOPS | 8x Tesla V100 16GB | 300 GB/s NVLink Hybrid Cube Mesh
2x Xeon | 7 TB RAID 0 | Quad IB 100Gbps, Dual 10GbE | 3U — 3200W
NEXT-GENERATION NVIDIA NVLINK FOR V100

300 GB/sec per GPU, 10x Faster than PCIe Gen3

NVLink for Tesla Pascal

NVLink for Tesla Volta

2 Rings

3 Rings
NVIDIA DGX-1
Customer Momentum
RIKEN SUCCESS STORY
Fujitsu and NVIDIA Build AI Supercomputer With 24 DGX-1s

CHALLENGES
Enterprises and research organizations embracing AI/DL
Needed to accelerated research in areas including medicine, manufacturing and healthcare
Conventional HPC architectures too costly and inefficient

SOLUTION
Partnered with Fujitsu for scale-out AI architecture built on DGX-1
24 DGX-1’s deliver 4 petaflops powering the RIKEN supercomputer
NVIDIA Cloud Services streamlines AI researcher workflow, helping accelerate RIKEN productivity

IMPACT
Accelerated real-world implementation of scale-out AI
Enables RIKEN team to take advantage of next-gen DL algorithms
Helping create future in which AI finds solutions to societal issues
<table>
<thead>
<tr>
<th>CHALLENGES</th>
<th>SOLUTION</th>
<th>IMPACT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinical data science center needed to apply ML to medicine</td>
<td>1&lt;sup&gt;st&lt;/sup&gt; medical institute in the world to leverage the DGX-1</td>
<td>New prostrate cancer pathology developed on DGX-1 in 6 months</td>
</tr>
<tr>
<td>Data volume requires immense computational capacity to process</td>
<td>Center for Clinical Data Sciences expands to partner hosp. (3X data)</td>
<td>AI/DL becomes critical tool in physician’s toolkit in 5-10 years</td>
</tr>
<tr>
<td>Immediate applications include radiology to improve accuracy, reduce variation</td>
<td>Deployment has grown to scale-out architecture with 4 DGX-1’s</td>
<td>Advancements in diagnostics, genomics and genetics</td>
</tr>
</tbody>
</table>
“Today’s practitioners have a barrage of data thrown at them — lab reports, MRIs, CAT scans, family health histories and more — which makes it incredibly difficult to make decisions. So, having technology that can aid them in this effort can be incredibly transformative.”

--Dr. Mark Michalski, CCDS Executive Director
The Pint-Sized Supercomputer That Companies Are Scrambling to Get

https://www.technologyreview.com/s/603075/the-pint-sized-supercomputer-that-companies-are-scrambling-to-get/

“The cost of renting enough servers on Amazon Web Services would surpass the system’s $129,000 price tag within a year.”

-Jackie Hunter, CEO, BenevolentAI

DGX-1

3x-4x

FASTER TRAINING
DGX STATION OVERVIEW
INTRODUCING NVIDIA DGX STATION

The Fastest Personal Supercomputer for Researchers and Data Scientists

- Revolutionary form factor - designed for the desk, whisper-quiet
- Start experimenting in hours, not weeks, powered by DGX Stack
- Productivity that goes from desk to data center to cloud
- Breakthrough performance and precision - powered by Volta
INTRODUCING NVIDIA DGX STATION
Groundbreaking AI - at your desk

The Personal AI Supercomputer for Researchers and Data Scientists

Key Features
1. 4 x NVIDIA Tesla V100 GPU
2. 2\textsuperscript{nd}-gen NVLink (4-way)
3. Water-cooled design
4. 3 x DisplayPort (4K resolution)
5. Intel Xeon E5-2698 20-core
6. 256GB DDR4 RAM
### At a Glance

<table>
<thead>
<tr>
<th>Feature</th>
<th>Specification</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>GPUs</strong></td>
<td>4x NVIDIA® Tesla® V100</td>
</tr>
<tr>
<td><strong>TFLOPS (GPU FP16)</strong></td>
<td>500</td>
</tr>
<tr>
<td><strong>GPU Memory</strong></td>
<td>16 GB per GPU</td>
</tr>
<tr>
<td><strong>NVIDIA Tensor Cores</strong></td>
<td>2,560 (total)</td>
</tr>
<tr>
<td><strong>NVIDIA CUDA Cores</strong></td>
<td>20,480 (total)</td>
</tr>
<tr>
<td><strong>CPU</strong></td>
<td>Intel Xeon E5-2698 v4 2.2 GHz (20-core)</td>
</tr>
<tr>
<td><strong>System Memory</strong></td>
<td>256 GB LRDIMM DDR4</td>
</tr>
<tr>
<td><strong>Storage</strong></td>
<td>Data: 3 x 1.92 TB SSD RAID 0 OS: 1 x 1.92 TB SSD</td>
</tr>
<tr>
<td><strong>Network</strong></td>
<td>Dual 10GBASE-T LAN (RJ45)</td>
</tr>
<tr>
<td><strong>Display</strong></td>
<td>3x DisplayPort, 4K Resolution</td>
</tr>
<tr>
<td><strong>Additional Ports</strong></td>
<td>2x eSATA, 2x USB 3.1, 4x USB 3.0</td>
</tr>
<tr>
<td><strong>Acoustics</strong></td>
<td>&lt; 35 dB</td>
</tr>
<tr>
<td><strong>Maximum Power Requirements</strong></td>
<td>1500 W</td>
</tr>
<tr>
<td><strong>Operating Temperature Range</strong></td>
<td>10 - 30 °C</td>
</tr>
<tr>
<td><strong>Software</strong></td>
<td>Ubuntu Desktop Linux OS</td>
</tr>
<tr>
<td></td>
<td>DGX Recommended GPU Driver</td>
</tr>
<tr>
<td></td>
<td>CUDA Toolkit</td>
</tr>
</tbody>
</table>

**DGX STATION SPECIFICATIONS**
3X FASTER THAN THE FASTEST WORKSTATIONS

Supercomputing performance at your desk

- Water-cooled performance - the only workstation built on 4 Tesla V100’s
- 3X the performance of today’s fastest GPU workstations
- with 30% faster training over non-DGX stack solutions
- 5X increase in I/O performance with 4-way NVLink vs. PCIe-connected GPU’s

500 TFLOPS
NVIDIA DGX
SOFTWARE STACK

Fully Integrated Software for Instant Productivity

Advantages:

- Instant productivity with NVIDIA optimized deep learning frameworks
  - Caffe, CNTK, MXNet, PyTorch, TensorFlow, Theano, and Torch
- Performance optimized across the entire stack
- Faster Time-to-Insight with pre-built, tested, and ready to run framework containers
- Flexibility to use different versions of libraries like libc, cuDNN in each framework container
THE POWER TO RUN MULTIPLE FRAMEWORKS AT ONCE
Container Images portable across new driver versions

## Containerized Applications

<table>
<thead>
<tr>
<th>Framework</th>
<th>Tuned SW</th>
<th>CUDA RT</th>
</tr>
</thead>
<tbody>
<tr>
<td>TensorFlow</td>
<td>TF Tuned SW</td>
<td>CUDA RT</td>
</tr>
<tr>
<td>Caffe2</td>
<td>Caffe2 Tuned SW</td>
<td>CUDA RT</td>
</tr>
<tr>
<td>Pytorch</td>
<td>Pytorch Tuned SW</td>
<td>CUDA RT</td>
</tr>
<tr>
<td>Other Frameworks and Apps</td>
<td>Tuned SW</td>
<td>CUDA RT</td>
</tr>
</tbody>
</table>

Linux Kernel + CUDA Driver

NVIDIA © DGX Systems
DL FROM DEVELOPMENT TO PRODUCTION
Accelerated Deep Learning Value with DGX Solutions

From Desk
- Procure DGX Station
- Install / Compile

Installed
- Fast Bring-up

Optimized
- Productive Experimentation
  - Experiment
  - Tune / Optimize

To Data Center or To Cloud
- Training at Scale
  - Deploy
  - Train
  - Insights

DGX Station
- DGX-1 / SATURNV / Cloud

Scaled
GE/AVITAS ENHANCE AI FOR ROBOTIC INSPECTION AND AUTOMATED DEFECT RECOGNITION

Powered by DGX-1 and DGX Station

HOW IT’S DONE

NVIDIA DGX Systems help Avitas Systems develop intelligent inspection services using autonomous and semi-autonomous robots, including drones, robotic crawlers, and autonomous underwater vehicles.

Avitas Systems uses deep learning to intelligently detect corrosion, leaks, fugitive emissions, and other defects imperceptible to the human eye with incredible accuracy that continually improves.

DGX Systems provide supercomputing power in the data center and the field, that combined with NVIDIA’s specialty-optimized deep learning frameworks, gives Avitas Systems an advantage over traditional inspection methods.
NVIDIA DGX SYSTEMS
Faster AI Innovation and Insight

The World’s First Portfolio of Purpose-Built AI Supercomputers

Get Started in AI - Faster
Effortless Productivity
Performance Without Compromise

For More Information: nvidia.com/dgx-systems