Making Multi-GPU Simulation Accessible to All with DGX A100 Appliances

Andy Lin | VP, Strategy & Innovation | Mark III Systems
GPU AND MULTI-GPU SIMULATIONS

Award-Winning Software Technology - Now a Whole Lot Faster!

Award-winning software technology

Parallelized using NVIDIA CUDA technology
  - Single GPU parallel since 2013
  - Multi-GPU capabilities added in 2021

Typical speedups up to 400x over serial on a multi-GPU workstation
WHY DOES SPEED MATTER?

1. Faster results and more cases simulated

Consider a 150x-200x speed-up:
   - Week long simulation in an hour

Increased exploration of research space enabling AI/ML

Fewer physical tests performed

New technologies brought to market faster, at lower cost.
WHY DOES SPEED MATTER?

2. Higher Fidelity Models

FCC regenerator simulation (~2012)
- Broke down the problem into 3 models
- Model 1: 2 days
- Model 2: 3 years
- Model 3: 64 days
WHY DOES SPEED MATTER?

2. Higher Fidelity Models

FCC regenerator simulation (~2012)

- Broke down the problem into 3 models
  - Model 1: 2 days
  - Model 2: 3 years
  - Model 3: 64 days

Comparable models today run in 12-48 hours

No modeler would construct the same model today!

- Larger model domain, finer spatial resolution, more/detailed physical models, complex chemical reaction mechanisms, etc.
WHY DOES SPEED MATTER?

3. Making the impossible, possible

For example, TRI’s technology development process commercialization of trash to jet fuel technology!

- A 1500x speed-up shown since 2009 (see S32115 from GTC-21 for details)
- Highly Commended by IChemE at 2020 Global Awards for Process Automation and Digitalization

Complexity in such models today was physically impossible to simulate in the past
Mark III Systems: CPFD & NVIDIA Elite Partner

25 Years NVIDIA Elite Partner
(AI/ML, HPC, Visualization, VDI,
Networking, Simulation, GPU computing)

Cross Industry AI & GPU Partnerships

HPC/AI & GPU Computing Showcase Demo Lab

HPC/AI Integration Center “Integrate and Ship Anywhere”
(Global HQ = Houston)

NVIDIA Rising Star Partner of the Year

Unique “Full Stack” Team
DGX A100: ONE SYSTEM FOR ALL AI INFRASTRUCTURE

Flexible AI infrastructure that adapts to the pace of enterprise

- World’s first 5 PFLOPS AI system, built on NVIDIA A100
- A100 delivers 20X performance vs prior generation
- One universal building block for the AI data center
- Analytics, training, inference - all on one platform
- Any workload on any node - any time

DGX A100: ONE SYSTEM FOR ALL AI INFRASTRUCTURE

AI Infrastructure Re-Imagined, Optimized, and Ready for Enterprise AI-at-Scale
DGX: IT'S MORE THAN A BOX

Full-Stack, Purpose-Built and Optimized for AI with Direct Access to Expertise

APPLIANCE OR INFRASTRUCTURE
Deploy as a standalone computing appliance or full DGX POD infrastructure solution

FULLY OPTIMIZED AI STACK
Continuous software engineering from NVIDIA that delivers performance that gets better over time

SOFTWARE AND TOOLS
Pre-trained models, scripts, and more from NGC enable out-of-box productivity, and certified AI workflow solutions enable MLOps

FLEXIBLE DEPLOYMENT
Deployed in DGX-Ready Data Centers or your own data centers, with options for CAPEX or OPEX with leasing or -aaS

INDUSTRY SOLUTIONS
Solutions built on DGX such as DRIVE Constellation for autonomous vehicles or Clara for healthcare
# NVIDIA DGX A100 640G

<table>
<thead>
<tr>
<th>Feature</th>
<th>Specification</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>GPUs</strong></td>
<td>8x NVIDIA A100 80GB GPUs</td>
</tr>
<tr>
<td><strong>GPU Memory</strong></td>
<td>640 GB total</td>
</tr>
<tr>
<td><strong>Peak performance</strong></td>
<td>5 petaFLOPS AI</td>
</tr>
<tr>
<td><strong>NVSwitches</strong></td>
<td>6</td>
</tr>
<tr>
<td><strong>System Power Usage</strong></td>
<td>6.5kW max</td>
</tr>
<tr>
<td><strong>CPU</strong></td>
<td>Dual AMD Rome 7742&lt;br&gt;128 cores total, 2.25 GHz (base), 3.4GHz (max boost)</td>
</tr>
<tr>
<td><strong>System Memory</strong></td>
<td>2TB</td>
</tr>
<tr>
<td><strong>Networking</strong></td>
<td>8x Single-Port Mellanox ConnectX-6 200Gb/s HDR InfiniBand (Compute Network)&lt;br&gt;2x Dual-Port Mellanox ConnectX-6 200Gb/s HDR InfiniBand (Storage Network also used for Eth*)</td>
</tr>
<tr>
<td><strong>Storage</strong></td>
<td>OS: 2x 1.92TB M.2 NVME drives&lt;br&gt;Internal Storage: 30TB (8x 3.84TB) U.2 NVME drives</td>
</tr>
<tr>
<td><strong>Software</strong></td>
<td>Ubuntu Linux OS (5.3+ kernel)</td>
</tr>
<tr>
<td><strong>System Weight</strong></td>
<td>271 lbs (123 kgs)</td>
</tr>
<tr>
<td><strong>Packaged System Weight</strong></td>
<td>360 lbs (163 kgs)</td>
</tr>
<tr>
<td><strong>Height</strong></td>
<td>6U</td>
</tr>
<tr>
<td><strong>Operating temp range</strong></td>
<td>5°C to 30°C (41°F to 86°F)</td>
</tr>
</tbody>
</table>

### Key Features
- **10x Mellanox CX-6 200 Gb/s Network Interface**<br>500 GB/sec Peak Bi-directional Bandwidth
- **Dual 64-core AMD Rome CPUs and 2 TB RAM**<br>3.2X More Cores to Power the Most Intensive AI Jobs
- **8x NVIDIA A100 GPUs with 640GB Total GPU Memory**
- **12 NVLinks/GPU**<br>600 GB/sec GPU-to-GPU Bi-directional Bandwidth
- **6x NVIDIA NVSwitches**
- **4.8 TB/sec Bi-directional Bandwidth**<br>2X More than Previous Generation NVSwitch
- **30TB Gen4 NVME SSD**
- **50 GB/sec Peak Bandwidth**<br>2X Faster than Gen3 NVME SSDs
BUILDING LEADING EDGE IN EVERY INDUSTRY

Thousands of Successful DGX Customer Deployments to Date
ADOPTED BY LEADING COMPANIES ACROSS INDUSTRIES

DGX Station Delivers AI Supercomputing and Simulation to More Teams, From Anywhere
Breaking the data center barrier

**A Supercomputer With Just Two Cables**

**No Data Center, No Problem!**
A fully functional AI system out-of-the-box, a whisper-quiet solution

**Work from Anywhere AI Appliance**
Plug into any standard wall socket, and access resources whether you are in the office, home office, or thousands of miles away

**Instant Productivity**
Unpack to up-and-running in under an hour, now with server-class remote management capabilities
CPFD Benchmarks on DGX A100

DGX A100 in action

**Barracuda Virtual Reactor 21.1**
Simulation Speed vs Problem Size using NVIDIA A100 GPUs

- **CPU Only**
- **1 GPU**
- **2 GPUs**
- **3 GPUs**
- **4 GPUs**

### SMALL SIMULATIONS

<table>
<thead>
<tr>
<th>Case</th>
<th>(FBR, coarse)</th>
<th>Simulation Speed (seconds/day)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>250,000 cells</td>
<td>11.4</td>
</tr>
<tr>
<td>2</td>
<td>105,000 cells</td>
<td>4.0</td>
</tr>
<tr>
<td>3</td>
<td>390,000 cells</td>
<td>1.5</td>
</tr>
<tr>
<td>4</td>
<td>820,000 cells</td>
<td>1.4</td>
</tr>
</tbody>
</table>

### LARGE SIMULATIONS

<table>
<thead>
<tr>
<th>Case</th>
<th>(FBR, coarse)</th>
<th>Simulation Speed (seconds/day)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>29.9 million cells</td>
<td>4.0</td>
</tr>
<tr>
<td>2</td>
<td>40.6 million cells</td>
<td>1.0</td>
</tr>
<tr>
<td>3</td>
<td>50.2 million cells</td>
<td>0.4</td>
</tr>
<tr>
<td>4</td>
<td>55.1 million cells</td>
<td>0.4</td>
</tr>
</tbody>
</table>
CPFD + DGX Station A100 - Options

Option #1: CPFD + DGX Station A100 (Acquire DGX Station & CPFD Integrated onsite)

Option #2: CPFD + DGX Station A100 (Pre-integrated & Shipped)

(Houston, TX)