HPC + AI OVERVIEW

Eric Kang 康勝閔 erkang@nvidia.com
Senior Manager, Solution Architect, NVIDIA
AGENDA

1. HPC + AI Overview - Eric Kang (15 Mins)

2. GPU acceleration principle and programing (HPC SDK, CUDA and OpenACC) - Ryan Jeng (30 Mins)

3. AI Principle and SDK - DL development framework - Jay Chen (15 Mins)

4. NVIDIA RAPIDS - Andrew Liu (15 Mins)
AI - A NEW INSTRUMENT FOR SCIENCE

HPC
- +40 years of algorithms based on first principles theory.
- Proven statistical models for accurate results

AI
- Improve predictive accuracy and faster response time.
- Previously unmanageable data sets

Dramatically Improves Accuracy and Time-to-Solution

- Commercially viable fusion energy
- Understanding cosmological dark energy and matter
- Clinically viable precision medicine
- Improvement and validation of the Standard Model of Physics
- Climate/weather forecasts with ultra-high fidelity
AI FOR SCIENCE
Transformative Tool To Accelerate The Pace of Scientific Innovation

IMPROVES ACCURACY
Enabling realization of full scientific potential

ACCELERATES TIME TO SOLUTION
Unlocking the use of science in exciting new ways
NVIDIA AI SUPERCOMPUTING DELIVERING SCIENTIFIC BREAKTHROUGHS

7 of 10 Gordon Bell Finalists are Using NVIDIA Platform

- **AI-Driven Multiscale Simulation**
  - Largest MD Simulation Ever

- **Docking Analysis**
  - 58x Full Pipeline Speedup

- **DeepMD-Kit**
  - 1,000x Speedup

- **Square Kilometre Array**
  - 250GB/s Data Processed End-to-end

Fusing AI And Data Analytics
NVIDIA CLARA DISCOVERY

**SEARCH**
- Chemical Compounds
  - O(10^6)

**GENOMICS**
- Clara Parabricks

**STRUCTURE**
- Biological Target
  - CryoSPARC, Relion
  - AlphaFold

**DOCKING**
- AutoDock
  - O(10^3)

**SIMULATION**
- Drug Candidate
  - Schrodinger
  - NAMD, VMD, OpenMM

**IMAGING**
- Clara Imaging

**NLP**
- Literature
  - Biomegatron, BioBERT

**RAPIDS**
- Biomegatron, BioBERT

**RAPIDS**
- Real World Data

**RAPIDS**
- Biomegatron, BioBERT
NVIDIA HPC + AI SUPERCOMPUTING

HPC

Deep Learning

Data Analytics

Energy Efficiency

11x Higher in 5 Years

Performance on Top HPC Apps

AI Training & Inference Leadership

#1

MLPerf

19x Better TCO

Data Analytics

10TB dataset

(SQL / ETL / ML / NLP)

World’s Greenest

26.2 GF/W

Power efficiency from the highest NVIDIA GPU-powered systems on Green500 lists. Kepler K20 - SC14, Pascal P100 - SC16, Volta V100 - SC18, Ampere A100 - SC20

Geometric Mean of application speedups vs. P100 : Benchmark Application: Amber [PME-Cellulose_NVE], Chroma [HMC], GROMACS [ADH Dodec], MILC [Apex Medium], NAMD [stmv_nve_cuda], PyTorch (BERT Large Fine Tuner), Quantum Espresso [AUSURF112-jR]; Random Forest FP32 [make_blobs (160000 x 64 : 10)]; TensorFlow [ResNet-50], VASP 6 [Si Huge], |GPU node: with dual-socket CPUs with 4x P100, V100, or A100 GPUs
AMAZING EXPANSION OF NVIDIA ECOSYSTEM

Apps for Every Industry Reaching Billions of Users

80 New SDKs

COMPLETE SOFTWARE STACK

GROWING ECOSYSTEM

CUDA Downloads in 2020
6M

GPU-Accelerated Applications
1,800

AI Startups
6,500


6,500

DeepStream 5.0

DeepStream

RTX DI

DeepStream 5.0

CUDA

CUDA-X

DLSS 2.1

RTX DI

Optix 7.2

HPC SDK 20.9

RAPIDS 0.16

Parabricks 3.5

DeepStream 5.0

NSIGHT 2020.5

cuDNN 8.03

TensorRT 7.2

CUDA 11.1

CUDA

NCCL 2.7.8

GPUDirect Storage

MAGNUM IO

TensorRT 7.2

cuDNN 8.03

HPC SDK 20.9

CUDA 11.1

CUDA-X

Back to Home
# NVIDIA DATACENTER PLATFORM

<table>
<thead>
<tr>
<th>BUSINESS APPLICATIONS</th>
<th>Customer Engagement</th>
<th>Patient Diagnostics</th>
<th>Fraud Detection</th>
<th>Quality Assurance</th>
<th>Industrial Automation</th>
<th>Precision Marketing</th>
<th>Molecular Simulations</th>
<th>--</th>
</tr>
</thead>
<tbody>
<tr>
<td>NGC</td>
<td>APPLICATION FRAMEWORKS</td>
<td>SMART CITY</td>
<td>CONVERSATIONAL AI</td>
<td>AUTONOMOUS VEHICLES</td>
<td>RECOMMENDATION SYSTEMS</td>
<td>HEALTHCARE</td>
<td>--</td>
<td>----</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Metropolis</td>
<td>Jarvis</td>
<td>Drive</td>
<td>Merlin</td>
<td>Clara</td>
<td>--</td>
<td>----</td>
</tr>
<tr>
<td></td>
<td>DEVELOPER TOOLKITS</td>
<td>ML &amp; DATA ANALYTICS</td>
<td>AI TRAINING &amp; INFERENCE</td>
<td>HIGH PERFORMANCE COMPUTING</td>
<td>RENDERING &amp; VISUALIZATION</td>
<td></td>
<td>--</td>
<td>----</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RAPIDS</td>
<td>TensorRT</td>
<td>NVIDIA HPC SDK</td>
<td>IndexX</td>
<td>OptiX</td>
<td></td>
<td>----</td>
</tr>
<tr>
<td></td>
<td>ACCELERATION LIBRARIES</td>
<td>COMPUTE</td>
<td>NETWORKING, STORAGE &amp; SECURITY</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>----</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CUDA-X</td>
<td>DOCA</td>
<td>MAGNUM IO</td>
<td></td>
<td></td>
<td></td>
<td>----</td>
</tr>
<tr>
<td>NVIDIA CERTIFIED</td>
<td>SERVERS &amp; CLOUD</td>
<td>DGX</td>
<td>EGX</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>----</td>
</tr>
<tr>
<td></td>
<td>VALIDATED SOLUTIONS</td>
<td>Purpose Built</td>
<td>Mainstream &amp; Edge</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>----</td>
</tr>
<tr>
<td></td>
<td>HARDWARE TECHNOLOGIES</td>
<td>GPU</td>
<td>NVSwitch</td>
<td>BlueField DPU</td>
<td>SMART NIC</td>
<td>Mellanox Switch</td>
<td></td>
<td>----</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SOFTWARE HUB</th>
<th>HELM</th>
<th>Certified Containers</th>
<th>Pre-trained Models</th>
<th>SDKs</th>
<th></th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>OPERATIONS</th>
<th>TRITON INFERENCE SERVER</th>
<th>FLEET COMMAND</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NVIDIA GPU Operator</td>
<td>Red Hat</td>
<td>----</td>
</tr>
<tr>
<td></td>
<td>vCenter</td>
<td>vmware</td>
<td>----</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>MONITORING</th>
<th>DCGM</th>
<th>Grafana</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DCMG</td>
<td>UFM</td>
<td>----</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>NETWORKING, STORAGE &amp; SECURITY</th>
<th>MAGNUM IO</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>RECOMMENDATION SYSTEMS</th>
<th>Merlin</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>HEALTHCARE</th>
<th>Clara</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>APPLICATION FRAMEWORKS</th>
<th>SMART CITY</th>
<th>CONVERSATIONAL AI</th>
<th>AUTONOMOUS VEHICLES</th>
<th>RECOMMENDATION SYSTEMS</th>
<th>HEALTHCARE</th>
<th>--</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Metropolis</td>
<td>Jarvis</td>
<td>Drive</td>
<td>Merlin</td>
<td>Clara</td>
<td>--</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>DEVELOPER TOOLKITS</th>
<th>ML &amp; DATA ANALYTICS</th>
<th>AI TRAINING &amp; INFERENCE</th>
<th>HIGH PERFORMANCE COMPUTING</th>
<th>RENDERING &amp; VISUALIZATION</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RAPIDS</td>
<td>TensorRT</td>
<td>NVIDIA HPC SDK</td>
<td>IndexX</td>
<td>OptiX</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ACCELERATION LIBRARIES</th>
<th>COMPUTE</th>
<th>NETWORKING, STORAGE &amp; SECURITY</th>
<th>DOCA</th>
<th>MAGNUM IO</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CUDA-X</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SERVERS &amp; CLOUD</th>
<th>DGX</th>
<th>EGX</th>
</tr>
</thead>
<tbody>
<tr>
<td>Purpose Built</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>HARDWARE TECHNOLOGIES</th>
<th>GPU</th>
<th>NVSwitch</th>
<th>BlueField DPU</th>
<th>SMART NIC</th>
<th>Mellanox Switch</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
NGC CATALOG - GPU-OPTIMIZED HPC & AI SOFTWARE

Easily Deploy Latest Software, Anywhere

NGC Catalog

SIMULATION
- AutoDock 4
- Chroma
- Folding @ Home
- LAMMPS

ARTIFICIAL INTELLIGENCE
- mxnet
- PyTorch
- TensorFlow
- Triton

VISUALIZATION
- IndeX
- OptiX
- ParaView
- VMD

x86 | Arm | POWER

DGX A100
DATA CENTER
EDGE
CLOUD
NVIDIA A100 80GB
Supercharging The World’s Highest Performing AI Supercomputing GPU

80GB HBM2e
For largest datasets and models

2TB/s +
World’s highest memory bandwidth to feed the world’s fastest GPU

3rd Gen Tensor Core

Multi-Instance GPU

3rd Gen NVLink
NVIDIA SELENENow Featuring NVIDIA DGX A100 640GB

4,480 A100 GPUs
560 DGX A100 system
850 Mellanox 200G HDR switches
14 PB of high-performance storage
2.8 EFLOPS of AI peak performance
63 PFLOPS HPL @ 24GF/W (Top500 #5 / Green500 #5)
DGX A100 STATION

SERVER-CLASS SOLUTION IN OFFICE-FRIENDLY FORM

Data Center Technology Outside the Data Center

First and only workstation with 4-way NVIDIA A100, NVLink and MIG

- Four A100 Tensor Core GPUs, 320 GB total HBM2E
- Multi-Instance GPU (MIG)
- 3rd generation NVLink
- 200 GB/s bi-directional bandwidth between any GPU pair, almost 3x compared to PCIe Gen4
- New Maintenance-free Refrigerant Cooling System

CPU and Memory
64-core AMD® EPYC® CPU, PCIe Gen4
512 GB system memory

Internal Storage
1.92 TB NVME M.2 SSD for OS
7.68TB NVME U.2 SSD for data cache

Connectivity
2x 10GbE (RJ45)
4x Mini DisplayPort for display out
Remote management 1GbE LAN port (RJ45)
加入 NVIDIA 開發者計畫

請掃描左邊 QR code 加入 NVIDIA 開發者計畫，活動結束前完成註冊，即可換取精美小禮物。

請向 NVIDIA 工作人員出示主旨為「Your application for the program NVIDIA Developer Program is approved」的 email。

developer.nvidia.com/developer-program