The Path from HPC to AI in NCHC

Ce-Kuen Shieh
General Director of NCHC
NCHC Milestones

1991
Taiwan’s first National level supercomputer Center

2003
NPO under NARLabs

2005
Tainan Office

2011
177 TF Windrider supercomputer

2017
1.325 PF TAIWANIA Peta scale HPC

1993
Hsinchu Headquarters

2004
TWAREN Services 10G

2008
Taichung Office

2016
100G Network Backbone

2018
Launch TAIWANIA Service & Start deploying AI/HPC

Certifications
✓ ISO 9001:2015
✓ ISO 27001:2013
✓ CSA STAR Level 2 Gold Award
✓ BS 10012
TWAREN-100G Backbone

TaiWan Advanced & REsearch Network

- **100G Bandwidth**
  - 12 regional networks
  - 94 universities & research institutes
  - 500 K. users

- **20G Bandwidth**
  - w/35 int’l research networks

Over 99.99% Network Availability

TWAREN Domestic Backbone

TWAREN International Connection
First Petascale supercomputer for open service in Taiwan

- Legacy HPC with 630 CPU nodes
- Performance: 1.325 Pflops
- Top500 Rank: 95 (Nov 2017)

Lanced 2018/07/01

Major users
- Universities
- Research Institutes

Major domains
- Science
- Engineering
Grand Plan in Taiwan

5 Strategies

- National Infrastructure for AI (NCHC NARLabs)
- AI Research Center (NTU, NCKU, NCTU, NTHU)
- Robot Makerspace (Science Parks)
- AI Chip Moonshot Program (MOST)
- (Universities & Research Organizations)
“National Infrastructure for AI” Strategy

- 4-year National-Level project
  - Build a National AI Infrastructure, including Compute, Storage and Network, for R&D
- Total Budget
  - US $ 150M
- Focus
  - Construct the TaiWan Computing Cloud (TWCC)
- Approach
  - Leverage Taiwan’s ICT superiority
- Benefit
  - Provide a common and sharable AI and Big Data Computing Platform across academia and industry to support AI Open Innovation Ecosystem and to accelerate AI research and industry
## National Infrastructure for AI

### Data Market
- Bio
- IoT
- Speech
- Image
- Video

### Software Platform
- Machine Learning
- Learned Neural Network
- Visualization
- Data Analytics
- Modal Simulation

### Application
- Precision Medicine
- Intelligent Machinery
- Technologies / Humanities
- Environment/ Disaster
- IoT
- Green Energy
- Fintech

### Cloud-based resource management, software and infrastructure
- TWAREN Backbone (100 Gbps)
- Storage (150 PB)
- CPU/GPU (10+ PF)
TaiWan Computing Cloud (TWCC)

Target for first 2 years budget (US$ 50 M)
TaiWan Computing Cloud (TWCC)

- Subsystem
  - AI/HPC
    - Deep Learning
    - HPC
  - Big Data Platform
    - VM
    - Big data analysis
  - Storage
    - Software Defined Storage
TaiWan Computing Cloud (TWCC)

• Vendor
  – A-Team
    • Quanta Computer, ASUS, Taiwan Mobile
  – Start deploying from April, 2018
TaiWan Computing Cloud (TWCC)

• Big Data Platform (250 nodes)
  – VMs
    • Hadoop
    • Spark
    • Hive
    • Hbase
    • ...
TaiWan Computing Cloud (TWCC)

- AI/HPC (9 Pflops)
  - Deep Learning
  - HPC

- Heterogeneous Storage System
  - Object/block storage
  - Parallel file system
  - Tape archive system
<table>
<thead>
<tr>
<th>Hardware - whole system</th>
<th>Software Environment</th>
</tr>
</thead>
<tbody>
<tr>
<td>• 252 nodes / 9072 CPU cores / 2016 GPUs</td>
<td>• Slurm / Kubernetes</td>
</tr>
<tr>
<td>• 193.5 TB memory</td>
<td>• Nvidia NGC Docker</td>
</tr>
<tr>
<td>• 10 PB storage</td>
<td>• Ceph</td>
</tr>
<tr>
<td>• EDR InfiniBand 100 Gbps</td>
<td>• Spectrum Scale (GPFS)</td>
</tr>
<tr>
<td>• 1.2 PUE (Warm Water Cooling)</td>
<td>• CentOS</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Hardware - single node</th>
<th>AI Framework</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Intel Xeon Gold CPU x 2</td>
<td>• Tensorflow</td>
</tr>
<tr>
<td>• Nvidia Tesla V100 w/32GB x 8</td>
<td>• Caffe / Caffe 2</td>
</tr>
<tr>
<td>• 768 GB memory</td>
<td>• PyTorch / Torch</td>
</tr>
<tr>
<td>• 240 GB SSD + 4TB NVMe</td>
<td>• ......and more</td>
</tr>
</tbody>
</table>
Software Architecture of TWCC

Cloud-Ready
- Self-service user portal and admin portal
- Multi-tenant cloud architecture
- VM and container-based provisioning
- System monitoring and management

End-to-End AI/HPC
- VM on OpenStack for data pre-processing
- Kubernetes with Docker for AI inferencing
- Kubernetes with NVidia-Docker for AI training
- Singularity for legacy HPC workload
<table>
<thead>
<tr>
<th>Application</th>
<th>Service</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smart City</td>
<td>Image Recognition</td>
</tr>
<tr>
<td>Disaster Mitigation</td>
<td>Text Mining</td>
</tr>
<tr>
<td>Industry 4.0</td>
<td>Natural Language Processing</td>
</tr>
<tr>
<td>Autonomous Car</td>
<td>Vehicle Speed Estimation</td>
</tr>
<tr>
<td>Biomedical Science</td>
<td>Bioinformatics Analysis</td>
</tr>
<tr>
<td>Robotics</td>
<td></td>
</tr>
</tbody>
</table>