Power and Energy Aware Computing with Tsubame 2.0 and Beyond

Satoshi Matsuoka
Tokyo Institute of Technology
2020 IT Carbon Footprint\(\approx\) Energy
(Original slide courtesy Horst Simon @ DoE-LBL)

“SMART 2020: Enabling the Low Carbon Economy in the Information Age”, The Climate Group

**Fig. 2.3 The global footprint by subsector**

<table>
<thead>
<tr>
<th>Emissions by geography</th>
<th>% of GtCO(_2)e</th>
</tr>
</thead>
<tbody>
<tr>
<td>2002</td>
<td>29</td>
</tr>
<tr>
<td>2007</td>
<td>37</td>
</tr>
<tr>
<td>2020</td>
<td>25</td>
</tr>
</tbody>
</table>

Telecoms  DC  PCs

**CAGR %**

<table>
<thead>
<tr>
<th>2002</th>
<th>2007</th>
<th>2020</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>7</td>
<td>5</td>
</tr>
</tbody>
</table>

* Printers were 11% of the total ICT footprint in 2002, 8% in 2007 and will be 12% in 2020.
†RoW = Rest of the World. (includes India, Brazil, South Africa, Indonesia and Egypt)

<table>
<thead>
<tr>
<th>140MW @ Tokyo Cloud Datacenter</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt; 100m diameter</td>
</tr>
<tr>
<td>9 stories 140,000 sq. ft.</td>
</tr>
</tbody>
</table>
OK, So We go Green Energy---
But It is Not that Simple

- Most “Green Energy” are Volatile

- 3/11 in Japan: even “non-volatile” power sometimes turns out to be volatile
  - Nuclear power plants knocked out
  - Fossil fuel plants takes time to restart
  - Result: Rolling Blackout Tokyo Area
3.11 Japan Earthquake

- Top 10 Supercomputers in Japan (96% perf. Share) all near Tokyo!
  - 18 of 26 Japanese SCs on the Top500

- All had to be stopped due to rolling blackouts
- Despite all the research, we were not ready for this...

### TOP500 Sublist Generator

\( R_{\text{max}} \) and \( R_{\text{peak}} \) values are in TFlops. For more details about other fields, check the TOP500 description.

26 entries found.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Site</th>
<th>System</th>
<th>Cores</th>
<th>( R_{\text{max}} )</th>
<th>( R_{\text{peak}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
So What Can We Do, As Computer Scientists?

- **Green Computing (Energy and Volatility)**
  - JST CREST Megascale Project [2003-2007]
  - JST CREST Ultra Low-Power HPC [2007-2012]
  - Tsubame2.0 “Green Supercomputer“ [2010-2014]
  - Green Supercomputing [2011-2015]

- **Using IT to Analyze and Optimize Energy**
  - Various Environmental and Device Simulations (on Tsubame2.0)
  - Supercomputing the Smart Grid - Post Petascale Graph Processing
JST-CREST “MegaScale” [2003-2007]

• “Embedded tech. meets HPC”
• Increased importance w/emphasis on mobile/ubiquitous computing
• Attain ~x2 with effective DVS

=> 2004 CTWatch J. (Matsuoka ed.)
• Can’t attain x10 tho...
  – Low hanging fruit achieved with simple DVS
  – Memory, NW, I/O... starts to dominate
  – Simple slow cores limited to weak scaling

– Conclusion: new devices, new architectures, cooling, apps need to be exploited and optimized as a whole
MegaProto [SC05] Computing Element

- V1: TM5900
  - 0.9GFlops@7W
  - L2C=512KB
  - 112.5MFlops/W
- V2: Effecion TM8800
  - 2.4GFlops@5W
  - 512MB DDR-SDRAM
  - 480MFlops/W

- 256MB SDRAM
  - (512M DDR in V2)
- 512KB flash
MegaProto Packaging Metrics

- **Goal: 300W/1U system**
  - CPU = 4-5W
  - Card = 10W → x17 = 170W
  - MB (Networks, interface, HDD) = 70W
    - Low power commodity fabric still problematic
  - Fans = ~20W
  - Power Supply Loss = 85% ~ 40W
  - **Total: ~300W**
    - (c.f. Xeon 3Ghz Dual 1U ~ 400W)
  - Current: \(.933 \text{ GF} \times 16 \div 300W = 50\text{MFlops/W}\) (15GigaFlops/1U)
  - **Next Version: 2.4 GF \times 16 \div 300W = 130\text{MFlops/W}\) (~40GigaFlops/1U)
    - x3 improvement in perf/power
MegaProto System

- Fabrication by IBM Japan (same fab team as BlueGene/L)
- 8 Units delivered, working
- 5 Units in our lab
MegaProto Integrated Network

- Low power, HP building block
- Reduce external switch requirements
MegaProto System Configuration

627GFlops@12.6KW per Rack
=> V2 1.61TFlops, 1.41TB/s MemBW, 1.3Gbps NetBW
The TSUBAME 1.0 “Supercomputing Grid Cluster”
Spring 2006

Unified IB network
Voltaire ISR9288 Infiniband 10Gbps
x2 (DDR next ver.)
~1310+50 Ports
~13.5Terabits/s (3Tbits bisection)

10Gbps x External Network

NEC SX-8i
(for porting)

Sun Galaxy 4 (Opteron Dual core 8-socket)
10480core/655Nodes
21.4Terabytes
50.4TeraFlops
OS Linux (SuSE 9, 10)
NAREGI Grid MW

“Fastest Supercomputer in Asia” 7th on the 27th
Top500@38.18TF

Storage
1.0 Petabyte (Sun “Thumper”)
0.1Petabyte (NEC iStore)
Lustre FS, NFS, WebDAV (over IP)
50GB/s aggregate I/O BW

ClearSpeed CSX600
SIMD accelerator
360 boards,
35TeraFlops(Current))
You know you have a problem when, ...

Turn on an electric fan at the temperature to 28°C in an air conditioned room.
Biggest Problem is Power…

<table>
<thead>
<tr>
<th>Machine</th>
<th>CPU Cores</th>
<th>Watts</th>
<th>Peak GFLOPS</th>
<th>Peak MFLOPS/Watt</th>
<th>Watts/CPU Core</th>
<th>Ratio c.f. TSUBAME</th>
</tr>
</thead>
<tbody>
<tr>
<td>TSUBAME (Opteron)</td>
<td>10480</td>
<td>800,000</td>
<td>50,400</td>
<td>63.00</td>
<td>76.34</td>
<td></td>
</tr>
<tr>
<td>TSUBAME 2006 (w/360CSs)</td>
<td>11,200</td>
<td>810,000</td>
<td>79,430</td>
<td>98.06</td>
<td>72.32</td>
<td></td>
</tr>
<tr>
<td>TSUBAME 2007 (w/648CSs)</td>
<td>11,776</td>
<td>820,000</td>
<td>102,200</td>
<td>124.63</td>
<td>69.63</td>
<td>1.00</td>
</tr>
<tr>
<td>Earth Simulator</td>
<td>5120</td>
<td>6,000,000</td>
<td>40,000</td>
<td>6.67</td>
<td>1171.88</td>
<td>0.05</td>
</tr>
<tr>
<td>ASCI Purple (LLNL)</td>
<td>12240</td>
<td>6,000,000</td>
<td>77,824</td>
<td>12.97</td>
<td>490.20</td>
<td>0.10</td>
</tr>
<tr>
<td>AIST Supercluster (Opteron)</td>
<td>3188</td>
<td>522,240</td>
<td>14400</td>
<td>27.57</td>
<td>163.81</td>
<td>0.22</td>
</tr>
<tr>
<td>LLNL BG/L (rack)</td>
<td>2048</td>
<td>25,000</td>
<td>5734.4</td>
<td>229.38</td>
<td>12.21</td>
<td>1.84</td>
</tr>
<tr>
<td>Next Gen BG/P (rack)</td>
<td>4096</td>
<td>30,000</td>
<td>16384</td>
<td>546.13</td>
<td>7.32</td>
<td>4.38</td>
</tr>
<tr>
<td>TSUBAME 2.0 (2010Q3/4)</td>
<td>160,000</td>
<td>810,000</td>
<td>1,024,000</td>
<td>1264.20</td>
<td>5.06</td>
<td>10.14</td>
</tr>
</tbody>
</table>

TSUBAME 2.0  x24 improvement in 4.5 years…? ➔ ~ x1000 over 10 years
**TSUBAME 1.2 Experimental Evolution (Oct. 2008)**

- **Voltaire ISR9288 Infiniband x8**
  - 10Gbps x2 ~1310+50 Ports
  - ~13.5Terabits/s
  - (3Tbits bisection)

- **NEC SX-8i**
  - 500GB 48disks

- **NEC SX-8i**
  - 10Gbps+External NW
  - Unified Infiniband network

- **87.1TF Linpack (The First GPU Supercomputer on Top500)**
  - [IPDPS10]

- **Storage**
  - 1.5 Petabyte (Sun x4500 x 60)
  - 0.1 Petabyte (NEC iStore)
  - Lustre FS, NFS, CIF, WebDAV (over IP)
  - 60GB/s aggregate I/O BW

- **10Gbps+External NW**
  - Unified Infiniband network

- **NEW Deploy: GCOE TSUBASA**
  - Harpertown-Xeon
  - 90Node 720CPU
  - 8.2TeraFlops

- **Storage**
  - Sun x4600 (16 Opteron Cores)
    - 32~128 GBytes/Node
    - 10480core/655Nodes
    - 21.4TeraBytes
    - 50.4TeraFlops
  - OS Linux (SuSE 9, 10)
  - NAREGI Grid MW

- **ClearSpeed CSX600**
  - SIMD accelerator
  - 648 boards, 52.2TeraFlops
  - SFP/DFP

- **New Production Experiment**
  - 170 Nvidia Tesla 1070, ~680 Tesla cards
  - High Performance in Many BW-Intensive Apps
  - 10% power increase over TSUBAME 1.0
モデルと実測の Bayes 的融合

- Bayes モデルと事前分布
- n 回実測後の事後予測分布

対象1（アルゴリズム1）
対象2（アルゴリズム2）

Auto-Tuning for Perf. & Power

Power Optimize using Novel Components in HPC

Power-Aware and Optimizable Applications
## Aggressive Power Saving in HPC

<table>
<thead>
<tr>
<th>Methodologies</th>
<th>Enterprise/Business Clouds</th>
<th>HPC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Server Consolidation</td>
<td>Good</td>
<td>NG!</td>
</tr>
<tr>
<td>DVFS (Differential Voltage/Frequency Scaling)</td>
<td>Good</td>
<td>Poor</td>
</tr>
<tr>
<td>New Devices</td>
<td>Poor (Cost &amp; Continuity)</td>
<td>Good</td>
</tr>
<tr>
<td>New Architecture</td>
<td>Poor (Cost &amp; Continuity)</td>
<td>Good</td>
</tr>
<tr>
<td>Novel Cooling</td>
<td>Limited (Cost &amp; Continuity)</td>
<td>Good (high thermal density)</td>
</tr>
</tbody>
</table>
E.g. CFD Applications

Comparing to TSUBAME Opteron CPU 1 core,

- Weather Prediction: x80 (10 GPU = 1000 CPUs)
- Lattice Boltzmann: ~ x100
- Phase Field Model: x170

When we assume the acceleration is typically x100,
1 GPU < 200W for our applications

100 GPU: 20 kW
10000 CPU core of TSUBAME 1.2: 1MW (1000kW)

X5~10 Low Power / Socket
Arithmetic INTENSITY: FLOP/Byte

FLOP = number of FP operation for applications
Byte = Byte number of memory access for applications

F = Peak Performance of floating point operation
B = Peak Memory Bandwidth

\[ \text{Performance} = \frac{\text{FLOP}}{\text{FLOP}/F + \text{Byte}/B + \alpha} \]

\[ = \frac{\text{FLOP}/\text{Byte}}{\text{FLOP}/\text{Byte} + F/B + \alpha} \]

\[ \alpha = \frac{\text{FLOP}}{\text{Byte}/B} \]
GPU vs. CPU Performance

Roofline model: Williams, Patterson 2008

Communication s of the ACM

FLOP/Byte = F/B

Performance [GFlops]

FLOP/Byte

708 GFlops

708 GFlops

x5–10 memory bound

x10–20 compute bound

GeForce GTX 285

Core i7 Extreme
From TSUBAME 1.2 to 2.0: From CPU Centric to GPU Centric Nodes for Scaling High Bandwidth in Network & I/O!!

CPU “Centric”
- DDR3 15-20GB STREAM per socket 2GB/core
- PCIe x16 4GB/s
- 40Gbps IB

Core Dataset
- IOH

GPU Vector
- GDDR5 150GB STREAM 3-6 GB per socket
- CPU Roles
  - OS
  - Services
  - Irregular Sparse
  - “Mixed” Algorithms

Core Dataset Vector Parallel
- PCIe x8 2GB/s x n
- IOH
- GPU
- 1GB
- Flash 200GB 400MB/s
- 54GB

“These are Isomorphic but much higher BW”
TSUBAME 2.0: A GPU-centric Green 2.4 Petaflops Supercomputer

Tsubame 2.0: "Tiny" footprint, very power efficient
- Floorspace less than 200m² (2,100 ft²)
- Top-class power efficient machine on the Green 500

TSUBAME 2.0
New Development

System
(42 Racks)
1408 GPU Compute Nodes,
34 Nehalem "Fat Memory" Nodes

Rack
(8 Node Chassis)

Chip
(CPU, GPU)

System

2.4 PFLOPS
80 TB

CPU (Westmere EP)
76.8 GFLOPS

GPUs (Tesla M2050)
515 GFLOPS
3 GB

GPU (Tesla M2050)
55 GB/103 GB

1.6 TFLOPS

6.7 TFLOPS
220 GB/412 GB

53.6 TFLOPS
1.7 TB/3.2 TB

Integrated by NEC Corporation
2.4 Petaflops, 1408 nodes
~50 compute racks + 6 switch racks
Two Rooms, Total 160m²
1.4MW (Max, Linpack), 0.48MW (Idle)
Tsubame2.0 Efficient Cooling Infrastructure

HP’s water-cooled rack
Completely closed racks with their own heat exchanger.
1.5 x width of normal rack+rear ext.
Cooling for high density deployments
35kW of cooling capacity single rack
  • Highest Rack Heat Density ever
  • 3000CFM Intake airflow with 7C chiller water
up to 2000 lbs of IT equipment
Uniform air flow across the front of the servers
Automatic door opening mechanism controlling both racks
Adjustable temperature set point
Removes 95% to 97% of heat inside racks
Polycarbonate front door reduces ambient noise considerably

~= Entire Earth Simulator (rack = 50TF)
TSUBAME2.0 World Rankings
(Nov. 2010 Announcement Green500!!!)

The Top 500 (Absolute Performance)
#1: ~2.5 PetaFlops: China Defense Univ. Dawning Tianhe 1-A
#2: 1.76 Petaflops: US ORNL Cray XT5 Jaguar
#3: 1.27 PetaFlops: China Shenzen SC Nebulae
#4: 1.19 PetaFlops: Japan Tokyo Tech. HP/NEC TSUBAME2.0
#5: 1.054 PetaFlops: US LLBL Cray XE6 Hopper
#~33 (#2 Japan): 0.191 Petaflops: JAEA Fujitsu

The Green 500 (Performance/Power Efficiency)
#1: 1684.20 : US IBM Research BG/Q Prototype (116)
#2: 958.35: Japan Tokyo Tech/HP/NEC Tsubame 2.0 (4)
#3: 933.06 : US NCSA Hybrid Cluster Prototype (403)
#4: 828.67: Japan Riken “K” Supercomputer Prototype (170)
#5-7: 773.38: Germany Julich etc.IBM QPACE SFB TR (207–209)
(#2+ 1448.03: Japan NAO Grape-DR Prototype) (383) (Added in Dec.)

TSUBAME2.0 “Greenest Production Supercomputer in the World”
Nov., 2010, June 2011 (two in a row!)
This certificate is in recognition of your organization’s achievements in reducing the environmental impact of high-performance computing.

GSIC Center, Tokyo Institute of Technology

Is recognized as the

Greenest Production Supercomputer in the World

on the world’s Green500 List of computer systems as of

November 2010

Wu-chun Feng, Co-Chair
Kirk Cameron, Co-Chair
Petaflops? Gigaflops/W?

Laptop: SONY Vaio type Z (VPCZ1)
CPU: Intel Core i7 620M (2.66GHz)
MEMORY: DDR3-1066 4GBx2
OS: Microsoft Windows 7 Ultimate 64bit
HPL: Intel(R) Optimized LINPACK Benchmark for Windows (10.2.6.015)
256GB HDD

18.1 GigaFlops Linpack
369 MegaFlops/W

Supercomputer: TSUBAME 2.0
CPU: 2714 Intel Westmere 2.93 Ghz
GPU: 4071 nVidia Fermi M2050
MEMORY: DDR3-1333 80TB + GDDR5 12TB
OS: SuSE Linux 11 + Windows HPC Server R2
HPL: Tokyo Tech Heterogeneous HPL 11PB Hierarchical Storage

1.192 PetaFlops Linpack
1043 MegaFlops/W

x66,000 faster
x3 power efficient
<<

x44,000 Data
## TSUBAME2.0 Power & PUE

<table>
<thead>
<tr>
<th></th>
<th>Power consumption by IT facility</th>
<th>Power consumption by Cooling facility</th>
<th>Total power consumption</th>
<th>PUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Typical operation</td>
<td>750kW</td>
<td>230kW</td>
<td>980kW</td>
<td>1.31</td>
</tr>
<tr>
<td>Under high utilization (GPU DGEMM)</td>
<td>1610kW</td>
<td>410kW</td>
<td>2020kW</td>
<td>1.25</td>
</tr>
</tbody>
</table>
The First Year of TSUBAME2.0

Operation started

Earthquake

Partial op.

“Peak shift” operation

“Short makespan” operation

Maintainance & Grand Ch.
## Operation History towards Shortage of Power

<table>
<thead>
<tr>
<th>Date</th>
<th>Operation</th>
<th># of nodes in operation</th>
</tr>
</thead>
<tbody>
<tr>
<td>2010/11/1</td>
<td>Operation started</td>
<td>S 300, G/V 480, H/X 420</td>
</tr>
<tr>
<td>2011/3/11</td>
<td>Earthquake</td>
<td></td>
</tr>
<tr>
<td>3/14</td>
<td>System shut down for rolling outage</td>
<td>0</td>
</tr>
<tr>
<td>3/17</td>
<td>Partially started, but stopped soon</td>
<td>S 160, G/V 100, H/X 150 -&gt; 0</td>
</tr>
<tr>
<td>3/2--4/6</td>
<td>Maintainance</td>
<td></td>
</tr>
<tr>
<td>4/6--4/7</td>
<td>Grand Challenge applications run</td>
<td></td>
</tr>
<tr>
<td>4/8--4/24</td>
<td>Partial operation</td>
<td>S 280, G/V 100</td>
</tr>
<tr>
<td>4/25—6/8,</td>
<td>“Peak shift” operation</td>
<td>S 300, G/V 480, + X 420 (only nighttime)</td>
</tr>
<tr>
<td>7/1—7/24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6/9—6/30</td>
<td>Full operation temporarily</td>
<td>S 300, G/V 480, H/X 420</td>
</tr>
<tr>
<td>7/25—9/25</td>
<td>“Short makespan” operation</td>
<td>S 300, G/V 280, H/X 420 Y 200 (makespan &lt;1h)</td>
</tr>
<tr>
<td>9/26—10/6</td>
<td>Grand Challenge applications run</td>
<td></td>
</tr>
<tr>
<td>10/6--</td>
<td>Full Operation</td>
<td>S 300, G/V 480, H/X 420</td>
</tr>
</tbody>
</table>
Realtime Measurement & Visualization of Power Consumption

- Now realtime power consumption information is available at
  [http://mon.g.gsic.titech.ac.jp/powermon/](http://mon.g.gsic.titech.ac.jp/powermon/)

*TSUBAME 2.0 Power Monitoring System*

---

Power is successfully reduced in daytime during peak shift operation
2011 MAGNITUDE 9
TOHOKU-OKI EARTHQUAKE

- Fatalities: 19,508
  - Strong shakings and devastating tsunamis
- Large source area
  - 500km x 200 km
  - Inner black rectangle
- Large FDM region required
  - 960km x 480km in horizontal
  - 240km in depth
  - Outer red rectangle
2011 MAGNITUDE 9 TOHOKU-OKI EARTHQUAKE
FDTD Simulation of Wave Propagation

- **Finite-Difference Time Domain** (Okamoto et al. 2010)
  - Topography, ocean layer, and heterogeneity
  - Grid size: 6400 x 3200 x 1600
  - Grid spacing: 150 m
  - Time interval: 0.005 s
  - **1000 GPUs of TSUBAME-2.0**
  - Preliminary source model

- **Visualization**
  - Vertical ground motion on land ocean bottom

Main part of the FDM region
2011 MAGNITUDE 9 TOHOKU-OKI EARTHQUAKE

FDTD Simulation of Wave Propagation

Main part of the FDM region
Power Consumption during 700-node Run

**Compute nodes (partial)**
- **903kW in total**
- **550kW for This app** (estimate from 540 nodes)

**Storage:**
- **72kW**

**Cooling:**
- **345kW at max** (shared by all jobs)

**Total:**
- **1320kW at max**

---

**2011 MAGNITUDE 9 TOHOKU-OKI EARTHQUAKE**

Power Consumption during 700-node Run

**Compute nodes (partial)**
- **903kW in total**
- **550kW for This app** (estimate from 540 nodes)

**Storage:**
- **72kW**

**Cooling:**
- **345kW at max** (shared by all jobs)

**Total:**
- **1320kW at max**

---

**TARO OKAMOTO**

**TOKYO INSTITUTE OF TECHNOLOGY**
Next Generation Numerical Weather Prediction [SC10]

Collaboration: Japan Meteorological Agency

Meso-scale Atmosphere Model:
Cloud Resolving Non-hydrostatic model
[Shimokawabe et. al. SC10 BSP Finalist]

ex. WRF (Weather Research and Forecast)

WSM5 (WRF Single Moment 5-tracer) Microphysics
Represents condensation, precipitation and thermodynamic effects of latent heat release
1% of lines of code, 25% of elapsed time
⇒ 20x boost in microphysics (1.2 - 1.3x overall improvement)

ASUCA: full GPU Implementation
developed by Japan Meteorological Agency

TSUBAME 2.0: 145 Tflops
World Record!!!
TSUBAME 2.0 Performance

Weak Scaling

145.0 Tflops
Single precision

76.1 Tflops
Double precision

Fermi core Tesla M2050
3990 GPUs

Previous WRF Record on ORN Jaguar
~ 50 TFLOPS (DFP)
x10 Socket-Socket
Power Consumption during Full TSUBAME2 Test with ASUCA

2011/04/08 2:18—2:26

![Graph showing power consumption during ASUCA run.]

- **Compute node**: $960kW$
- **Storage**: $78kW$
- **Cooling**: $270kW$ max
- **Total**: $1308kW$ max
NAMD is a parallel molecular dynamics code developed at University of Illinois.

This evaluation is result of an interdisciplinary collaboration between UIUC and Tokyo Tech.

The 100-million-atom benchmark in this work was assembled by replicating a million-atom satellite tobacco mosaic virus (STMV) simulation on a 5x5x4 grid.

One STMV (Satellite Tobacco Mosaic Virus) includes 1,066,628 atoms.
100-million-atom MD Simulation
M. Sekijima (Tokyo Tech), Jim Phillips (UIUC)

Performance Evaluation

<table>
<thead>
<tr>
<th></th>
<th>8 nodes</th>
<th>32 nodes</th>
<th>64 nodes</th>
<th>128 nodes</th>
<th>256 nodes</th>
<th>512 nodes</th>
<th>700 nodes</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPU 12 cores</td>
<td>0.00</td>
<td>0.02</td>
<td>0.03</td>
<td>0.06</td>
<td>0.11</td>
<td>0.21</td>
<td>0.28</td>
</tr>
<tr>
<td>CPU 12 cores + 1 GPU</td>
<td>N/A</td>
<td>N/A</td>
<td>0.06</td>
<td>0.15</td>
<td>0.31</td>
<td>0.60</td>
<td>0.78</td>
</tr>
<tr>
<td>CPU 12 cores + 2 GPUs</td>
<td>N/A</td>
<td>N/A</td>
<td>0.14</td>
<td>0.26</td>
<td>0.50</td>
<td>0.91</td>
<td>1.21</td>
</tr>
<tr>
<td>CPU 12 cores + 3 GPUs</td>
<td>N/A</td>
<td>N/A</td>
<td>0.16</td>
<td>0.31</td>
<td>0.58</td>
<td>1.00</td>
<td>1.32</td>
</tr>
</tbody>
</table>
100-million-atom MD Simulation

M. Sekijima (Tokyo Tech), Jim Phillips (UIUC)
Power Consumption during 700-node Run

Compute node:
- $1115\text{kW}$ in total
- $706\text{kW}$ for This app
  (estimate from 540 nodes)

Storage:
- $72\text{kW}$

Cooling:
- $340\text{kW}$ max
  (shared by all jobs)

Total:
- $1527\text{kW}$ max
Isotropic turbulence

**Pseudo Spectral Method**

(2/3 dealiasing)

\[ \text{Re} = 500 \]
\[ N = 2048^3 \]

**Vortex Particle Method**

(Reinitialized CSM)

\[ \text{Re} = 500 \]
\[ N = 2048^3 \]

8 billion particles
Petaflops scale turbulence simulation on TSUBAME 2.0

Weak Scaling

Wall clock time

Parallel efficiency

R. Yokota (KAUST), L. A. Barba (Boston Univ), T. Narumi (Univ of Electro Communications), K. Yasuoka (Keio Univ)
Petaflops scale turbulence simulation on TSUBAME 2.0

Present work

Rahimian et al. (2010 Gordon Bell)

64 billion in 100 seconds
1.0 PFlops

90 billion in 300 seconds
0.7 PFlops

R. Yokota (KAUST) , L. A. Barba (Boston Univ), T. Narumi (Univ of Electro Communications), K. Yasuoka (Keio Univ)
Petaflops scale turbulence simulation on TSUBAME 2.0

Power Usage during Full System Test

- **Compute node:** 1190kW
- **Storage:** 72kW
- **Cooling:** 240kW
- **Total:** 1502kW

2011/10/4 5:00—6:00

R. Yokota (KAUST), L. A. Barba (Boston Univ), T. Narumi (Univ of Electro Communications), K. Yasuoka (Keio Univ)
Graph500 is a new benchmark that ranks supercomputers by executing a large-scale graph search problem.

The benchmark is ranked by so-called **TEPS** (*Traversed Edges Per Second*) that measures the number of edges to be traversed per second by searching all the reachable vertices from one arbitrary vertex with each team’s optimized BFS (Breadth-First Search) algorithm.

**Kronecker graph**

A: 0.57, B: 0.19, C: 0.19, D: 0.05

**G₄ adjacency matrix**

Toyotaro Suzumura, Koji Ueno, Tokyo Institute of Technology
Highly Scalable Graph Search Method for the Graph500 Benchmark

• Our early study reveals that the provided reference implementations are not scalable in a large-scale distributed environment.

• We propose an optimized method based on 2D based partitioning and other various optimization methods such as communication compression and vertex sorting.

• Our optimized implementation can solve BFS (Breadth First Search) of large-scale graph with $2^{36}$ (68.7 billion) vertices and $2^{40}$ (1.1 trillion) edges for 10.58 seconds with 1366 nodes and 16392 CPU cores on TSUBAME 2.0

• This record corresponds to 103.9 GE/s (TEPS)

Vertex Sorting by utilizing the scale-free nature of the Kronecker Graph

2D Partitioning Optimization

Performance Comparison with Reference Implementations (simple, replicated-csr and replicated-csc) and Scale 24 per 1 node

Performance of Our Optimized Implementation with Scale 26 per 1 node

Toyotaro Suzumura, Koji Ueno, Tokyo Institute of Technology
Power Consumption during Graph500 Run on TSUBAME 2.0

Compute node: 902kW
Storage: 75kW
Cooling: 346kW max
Total: 1323kW max

2011/10/4 18:00—22:00
Background

Mechanical Structure

Material Microstructure

Improvement of fuel efficiency by reducing the weight of transportation and mechanical structures

Developing lightweight strengthening material by controlling microstructure

Low-carbon society

Dendritic Growth
Impact of Peta-scale Simulation on Material Science

Previous Research

2D

3D simple shape

Single dendrite

Peta-scale Simulation

✓ GPU-rich Supercomputer
✓ Optimization for Peta-scale computing

Distribution of multiple dendrites is important for design of solidified products.

Scientific meaningful 3D simulation
Observation and Simulation

Observation:
X-ray imaging of Solidification of a binary alloy at Spring-8 in Japan by Prof. Yasuda (Osaka University in Japan)

Our Simulation:
Phase-field simulation using mesh size of 4096 x 128 x 4096 on TSUBAME 2.0
Weak scaling results on TSUBAME 2.0

- **4096 x 6400 x 12800**
- **4000 (40 x 100) GPUs**
- **16,000 CPU cores**

**Methods**:
- **GPU-Only** (No overlapping)
- **Hybrid-YZ** (y,z boundary by CPU)
- **Hybrid-Y** (y boundary by CPU)

**Performance [TFlops]**

- **Hybrid-Y method**
  - 2.0000045 PFlops
  - GPU: 1.975 PFlops
  - CPU: 24.69 TFlops
  - Efficiency 44.5%
    - (2.000 PFlops / 4.497 PFlops)

**Parameters**:
- **Mesh size**: 4096 x 160 x 128 / GPU
- **NVIDIA Tesla M2050 card / Intel Xeon X5670 2.93 GHz on TSUBAME 2.0**
Power consumption and Efficiency

- The power consumption by application executions on TSUBAME 2.0 is measured in detail.
- Our phase-field simulation (real application)
  - 2.000 PFlops (single precision) 2PFlops-Simulation
  - Performance to the peak: 44.5%
  - Green computing: 1468 MFlops/W (Total: 1729kW)

We obtained the simulation results by small electric power consumption.

Ref. Linpack benchmark
  - 1.192 PFlops (DP)
  - Efficiency 52.1%
  - 827.8 MFlops/W
Power Consumption during 2.0PFlops Phase-Field Run

Compute node: 1362kW
Storage: 73kW
Cooling: 294kW at max
Total: 1729kW at max

2011/10/5 2:00—3:00
MUPHY: Multiphysics simulation of blood flow
(Melchionna, Bernaschi et al.)

Combined Lattice-Boltzmann (LB) simulation for plasma and Molecular Dynamics (MD) for Red Blood Cells

Realistic geometry (from CAT scan)

Two-levels of parallelism: CUDA (on GPU) + MPI

- 1 Billion mesh node for LB component
- 100 Million RBCs

Fluid: Blood plasma

Lattice Boltzmann

Irregular mesh is divided by using PT-SCOTCH tool, considering cutoff distance

Body: Red blood cell

Extended MD

Red blood cells (RBCs) are represented as ellipsoidal particles

Multiphysics simulation with MUPHY software
Cluster of Nvidia M2050 GPUs connected by QDR Infiniband. Scaling study up to 512 nodes (each node has 3 GPUs). Very fast parallel I/O (read 100 GB in ~10 sec)

<table>
<thead>
<tr>
<th>GPUs</th>
<th>Time (s)</th>
<th>Efficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td>256</td>
<td>0.07616</td>
<td>N.A.</td>
</tr>
<tr>
<td>512</td>
<td>0.03852</td>
<td>98.86 %</td>
</tr>
<tr>
<td>1,024</td>
<td>0.01995</td>
<td>95.37 %</td>
</tr>
<tr>
<td>1,536</td>
<td>0.01343</td>
<td>94.43 %</td>
</tr>
</tbody>
</table>

Lattice Boltzmann Scaling (time per step)

LB kernel: 1 GPU ~200 BG/P cores
1536 GPUs equivalent to full BlueGene/P

<table>
<thead>
<tr>
<th>GPUs</th>
<th>Time (s)</th>
<th>Efficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td>256</td>
<td>0.44453</td>
<td>N.A.</td>
</tr>
<tr>
<td>512</td>
<td>0.25601</td>
<td>86.82 %</td>
</tr>
<tr>
<td>1,024</td>
<td>0.14062</td>
<td>79.03 %</td>
</tr>
</tbody>
</table>

Lattice Boltzmann + Cell Dynamics Scaling (time per step)

Time to completion on stationary flow: 23 minutes

New run on FULL TSUBAME2.0 (4000 GPUs) just completed with an improved algorithm, exhibiting petascale performance(!!)
Results on Tsubame2 Supercomputer (2) : Using 4,000 GPUs

Strong Scaling Results

Elapsed time per timestep for 1G mesh nodes and 450M RBCs (log scale)

Parallel efficiency for 110, 220, 450M RBCs

~80% with 4K GPUs

Speeds per Component

0.6PFlops with 4,000GPUs for 1G mesh nodes, 450M RBCs

A complete heartbeat at microsecond resolution can be simulated in 48hours
## TSUBAME2.0 Power Consumption with Petascale Applications

<table>
<thead>
<tr>
<th></th>
<th>Compute nodes &amp; SW (kW)</th>
<th>Storage (kW)</th>
<th>Cooling (kW)</th>
<th>Total (kW)</th>
<th>Cooling/Togal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Typical operation</td>
<td>750</td>
<td>72</td>
<td>230</td>
<td>980</td>
<td>23.5%</td>
</tr>
<tr>
<td>Earthquake (700nodes)</td>
<td>550/903</td>
<td>72</td>
<td>345</td>
<td>1320</td>
<td>26.1%</td>
</tr>
<tr>
<td>MD (700nodes)</td>
<td>706/1115</td>
<td>72</td>
<td>340</td>
<td>1527</td>
<td>22.3%</td>
</tr>
<tr>
<td>ASUCA</td>
<td>960</td>
<td>78</td>
<td>270</td>
<td>1308</td>
<td>20.6%</td>
</tr>
<tr>
<td>Turblence</td>
<td>1190</td>
<td>72</td>
<td>240</td>
<td>1502</td>
<td>16.0%</td>
</tr>
<tr>
<td>Graph500</td>
<td>902</td>
<td>75</td>
<td>346</td>
<td>1323</td>
<td>26.2%</td>
</tr>
<tr>
<td>Phase-field</td>
<td>1362</td>
<td>73</td>
<td>294</td>
<td>1729</td>
<td>17.0%</td>
</tr>
<tr>
<td>GPU DGEMM</td>
<td>1538</td>
<td>72</td>
<td>410</td>
<td>2020</td>
<td>20.3%</td>
</tr>
</tbody>
</table>
モデルと実測の Bayes 的融合

\[
\begin{align*}
&\text{Bayes モデルと事前分布} \\
&y_i \sim N(\mu_i, \sigma_i^2), \\
&\mu_i \mid \beta, \sigma_i^2 \sim N(\alpha_i \beta, \sigma_i^2 / \kappa_i) \\
&\sigma_i^2 \sim \text{Inv-Chi}(\nu_i, \sigma_i^2)
\end{align*}
\]

\[
\begin{align*}
&\text{事後予測分布} \\
&y_i \mid (Y_{\text{后}}, \alpha_i, \beta, \kappa_i) \\
&\sim N(\mu_{i, \text{post}}, \sigma_{i, \text{post}}^2), \\
&\mu_{i, \text{post}} = \frac{\nu_i y_i + \nu_0 \mu_i}{\nu_i + \nu_0}, \\
&\sigma_{i, \text{post}}^2 = \frac{1}{\nu_i + \nu_0}\left[\nu_i (y_i - \mu_i)^2 + \nu_0 (\mu_{\text{post}} - \mu_i)^2 + \frac{\nu_i \nu_0 \sigma^2}{\kappa_i}ight]
\end{align*}
\]

Auto-Tuning for Perf. & Power

Power Optimize using Novel Components in HPC

Power-Aware and Optimizable Applications
Precise GPU Power Measurement

Analysis of lines of PCI Express riser card and extraction of power supply lines

<table>
<thead>
<tr>
<th>Line #</th>
<th>Description</th>
<th>Line #</th>
<th>Description</th>
<th>Line #</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>GND</td>
<td>13</td>
<td>GND</td>
<td>25</td>
<td>JTAG3</td>
</tr>
<tr>
<td>2</td>
<td>GND</td>
<td>14</td>
<td>PRSNT#1</td>
<td>26</td>
<td>JTAG4</td>
</tr>
<tr>
<td>3</td>
<td>GND</td>
<td>15</td>
<td>12V</td>
<td>27</td>
<td>JTAG5</td>
</tr>
<tr>
<td>4</td>
<td>GND</td>
<td>16</td>
<td>12V</td>
<td>28</td>
<td>3.3V</td>
</tr>
<tr>
<td>5</td>
<td>GND</td>
<td>17</td>
<td>12V</td>
<td>29</td>
<td>3.3V</td>
</tr>
<tr>
<td>6</td>
<td>GND</td>
<td>18</td>
<td>12V</td>
<td>30</td>
<td>3.3V</td>
</tr>
<tr>
<td>7</td>
<td>GND</td>
<td>19</td>
<td>12V</td>
<td>31</td>
<td>3.3V</td>
</tr>
<tr>
<td>8</td>
<td>GND</td>
<td>20</td>
<td>SMCLK</td>
<td>32</td>
<td>JTAG1</td>
</tr>
<tr>
<td>9</td>
<td>GND</td>
<td>21</td>
<td>12V(B3)</td>
<td>33</td>
<td>PWRGD</td>
</tr>
<tr>
<td>10</td>
<td>GND</td>
<td>22</td>
<td>SMDAT</td>
<td>34</td>
<td>3.3Vaux</td>
</tr>
<tr>
<td>11</td>
<td>GND</td>
<td>23</td>
<td>GND</td>
<td>35</td>
<td>WAKE#</td>
</tr>
<tr>
<td>12</td>
<td>GND</td>
<td>24</td>
<td>JTAG2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Automatic identification of measurement sequences and high precision power measurement with marker kernels

Correction formulae dep. probe and PSU

Precise measurements

Y = 38.6 + 1.15 X – 0.000637 (X – 100) (X – 200) ± 0.9

Ubiquitous Probe by Kuroda Team

Analysis of lines of PCI Express riser card and extraction of power supply lines

High: CudaMemcopy

Low: Comp w/ 1 thread

Correction formulae dep. probe and PSU

Precise measurements

Wireless measurements

Y = 38.6 + 1.15 X – 0.000637 (X – 100) (X – 200) ± 0.9
Measuring GPU Power Consumption

• Two power sources
  – Via PCI Express: < 75 W
  – Direct inputs from PSU: < 240 W

• Uses current clamp sensors around the power inputs

Precise and Accurate Measurement with Current Probes

Attaches current sensors to two power lines in PCIe

Direct power inputs from PSU

A/D converter

NI PCIe-6259

Measurement PC

Reads currents at 100 us interval
Statistical Power Modeling of GPUs

[IEEE IGCC10]

- Estimation of GPU power consumption GPU statistically
- Linear regression model using performance counters as explanatory variables

\[
p = \sum_{i=1}^{n} \alpha_i c_i + \beta
\]

- Prevents overtraining by ridge regression
- Determines optimal parameters by cross fitting

GPU performance counters

Average power consumption

Power meter with high resolution

High accuracy (Avg Err 4.7%)

Accurate even with DVFS

Future: Model-based power opt.

Linear model shows sufficient accuracy

Possibility of optimization of Exascale systems with O(10^8) processors
Optimization Effect Predictors for GPUs

Optimization effects heavily depend on GPU architecture and generation

Estimate effects before making efforts

Speed-up by “shared memory” optimization

Future work:
Estimating improvement of energy consumption by integrating power model

Comparing measured speed up and estimation
Low Power Scheduling in GPU Clusters

[IEEE IPDPS-HPPAC 09]

Objective: Optimize CPU/GPU Heterogeneous

- Optimally schedule mixed sets of jobs executable on either CPU or GPU but w/different performance & power
- Assume GPU accel. factor (%) known

30% Improvement Energy-Delay Product

TODO: More realistic environment

- Different app. power profile
- PCI bus vs. memory conflict
  - GPU applications slow down by 10% or more when co-scheduled with memory-intensive CPU app.
Intelligent Task Scheduling for GPU clusters - considering conflicts on memory bus and PCI-E bus -

4 processes share a node.

Assuming the following information of each job is known
• Execution time (at standalone case)
• Total memory access (from performance counter)
• Total PCI-E transfer (from CUDA profiler)
Actual execution time is estimated by our model.

We integrated this model into ECT (Earliest Complete Time) scheduling.

Energy reduction of up to 2.3%
ULP-HPC Research Example: Next-Gen ULP Memory System ([IEEE IPDPS-HPPAC 08])

- CPUs
  - L1 cache
  - L2 cache

MAIN MEMORY

- Non-Volatile • Low Power Fast Device
  - MRAM

- Decreased DRAM Requirement
  - DRAM

SWAP

- Non-Volatile • Low Power Medium-Spee Device for fast SWAP
  - FLASH

HDD
Performance vs. Power Consumption

Application: NAS CG (class B)
Standard system: DRAM 1GB
Proposed system:
(MRAM 128MB + DRAM 192MB)

Execution overhead: x 1.6
Memory Energy Consumption: reduced to 28%
Ultra-Low-Power HPC Interconnects

- **Objective:** Power efficiency of interconnects is improved by 100 times within 10 years
  - Using a large number of small switches
  - No more expensive monster-switches (1,000-port switch)

Reference

Traffic decreases (Links becomes off)
Research Outcome

Before

Monster switch (312 port@1)

After (ULP-HPC)

Small switch (48port@8)

SuperNova cluster (225 hosts, 93rd supercomputer on Nov, 2003)

**Power:** using small switches (-87%) × cyclic topology (+60%) × On/Off link regulation (-37%) ≈ -87% (1/8)

**Cost:** using small switches (-98%) × cyclic topology (+60%) × On/Off link regulation (±0%) ≈ -97% (1/30)

NII Group (Michihiro Koibuchi (koibuchi@nii.ac.jp), Masato Yoshimi (myosimi@mail.doshisha.ac.jp))
Green Technologies in Production in Tsubame2.0---ULPHPC Feedback

- Extensive Many-Core (GPU)-Multi-Core (CPU) Architecture
- Power efficient devices: 3000 SSDs (I/O), Integrated Fiber Optics, 92% efficient PSU, Fans and other LP devices
- Efficient hybrid Air-Liquid Cooling (HP-MCS Racks, PUE < 1.3)
- Extensive Thermal and Power Sensing (30,000 thermal sensors, power sensors / node & rack, etc.)
- Proactive Node/Rack Power Capping (used constantly)
- Power efficient software stack (DVFS etc.) and GPU-apps
- “Green” Peak-Shifting Scheduling- Dealing with Volatility (new!)
Dealing with Further Volatility and Power Requirements

- Up to x3 load dependent power variance
- Matching the power load to variance of the power infrastructure
DoE Exascale Parameters
x1000 power efficiency in 10 years

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>System peak</td>
<td>2 PetaFlops</td>
<td>100-200 PetaFlops</td>
<td>1 ExaFlop</td>
</tr>
<tr>
<td>Power</td>
<td>Jaguar 6 MW</td>
<td>TSUBAME 1.3 MW</td>
<td>15 MW</td>
</tr>
<tr>
<td>System Memory</td>
<td>0.3PB</td>
<td>0.1PB</td>
<td>5 PB</td>
</tr>
<tr>
<td>Node Perf</td>
<td>125GF</td>
<td>1.6TF</td>
<td>0.5TF</td>
</tr>
<tr>
<td>Node Mem BW</td>
<td>25GB/s</td>
<td>0.5TB/s</td>
<td>0.1TB/s</td>
</tr>
<tr>
<td>Node Concurrency</td>
<td>12</td>
<td>O(1000)</td>
<td>O(100)</td>
</tr>
<tr>
<td>#Nodes</td>
<td>18,700</td>
<td>1442</td>
<td>50,000</td>
</tr>
<tr>
<td>Total Node Interconnect BW</td>
<td>1.5GB/s</td>
<td>8GB/s</td>
<td>20GB/s</td>
</tr>
<tr>
<td>MTTI</td>
<td>O(days)</td>
<td>O(1 day)</td>
<td>O(1 day)</td>
</tr>
<tr>
<td>Machine</td>
<td>Power (incl. cooling)</td>
<td>Linpack Perf (PF)</td>
<td>Linpack Mflops/W</td>
</tr>
<tr>
<td>----------------------</td>
<td>-----------------------</td>
<td>-------------------</td>
<td>------------------</td>
</tr>
<tr>
<td>Tsubame1.0 (2006Q1)</td>
<td>1.8MW</td>
<td>0.038</td>
<td>21</td>
</tr>
<tr>
<td>ORNL Jaguar (2009Q4)</td>
<td>~9MW</td>
<td>1.76</td>
<td>196</td>
</tr>
<tr>
<td>Tsubame2.0 (2010Q4)</td>
<td>1.8MW</td>
<td>1.2</td>
<td>667</td>
</tr>
<tr>
<td>K Computer (2011Q2)</td>
<td>~16MW</td>
<td>10</td>
<td>625</td>
</tr>
<tr>
<td>BlueGene/Q (2012Q1)</td>
<td>~12MW?</td>
<td>17</td>
<td>1417</td>
</tr>
<tr>
<td>Tsubame3.0 (2015Q1)</td>
<td>1.8MW</td>
<td>20</td>
<td>11000</td>
</tr>
<tr>
<td>EXA (2018Q4)?</td>
<td>20MW</td>
<td>1000</td>
<td>50000</td>
</tr>
</tbody>
</table>