Delivering Optimized Platforms for Deep Learning and HPC Workloads
How Big is the Opportunity

AI Is Growing Exponentially

A CAGR of 57% over the period of 2016 - 2025

CAGR: Compound annual growth rate

Machine Learning & Big Data & Computing Capacity

Diverse Applications in the Market

AI related Solutions Development

(Source: Tractica)
AI: Is Here, and Everywhere

Research: Image recognition

Semiconductor: Image recognition

Social Media: Video/image/speech recognition

Retail: Data analysis and decision making

Everywhere
JOURNEY STARTS HERE 2008

1st Non-Blocking PCI-E Gen2 GPU Servers 2009

Enable GPU Solution in Blade 2011

Enable 3x GPU Solution per Blade & upto 12x GPU in 4U 2013

1st 4x GPU in 1U Server 2015

NVLINK and Single Complex Root 2016

Here&Now (NVLINK 2.0) Supported
GPGPU Solutions with Technologies Implementation

Certified Optimized Complete Solutions

Single Complex Root Topology
Hybrid Cube Mesh Topology

Integrated with Latest Technology…
Latency Optimized Solution

NO PCI-E BOTTLE Necks Constrain…
GPU RDMA Direct Supports with Optimized Latency Designs
With NVLINK Architecture Design-In

Green Computing Architecture
Finest Thermal and Power Delivery Platforms
Offering High-Dense and Scalable Solutions
Supermicro adopts quickly **NVLINK Technology** into Deep-Learning Platform
We Can Support up to **300GB/s** Total Bandwidth

Two Optimized NVLINK Solutions

1U in 4U

Scalability

Master Compute

Multi- Cooling Options*

*Depends on Project
Support Up to 165W TDP
**Scalability** Compute

Single Layer – **Thermal Design**

Best Option in **Scale-Out**

**Performance/ 4U**

**Low Initial Start Up Cost**

**Balanced Performance/ Cost**

Support Up to 205W TDP
**Master Purpose** Compute

Double Layer – **Thermal Design**

Best **Scale-Up Option**

**In Memory**

**PCI-E IO Expansion**

**Storages**

**Integrated with High Eff. PSU**
Benchmark - Performance Results

1, 2, 4, 8, 16, and 32 V100 SXM2 Configurations

We have validated with Three Frameworks

CAFFE2
- 1x V100 SXM2: 7%
- 8x V100 SXM2: 3%

RESNET 50
- 8x V100 SXM2: 11%

INCEPTION V3
- 8x V100 SXM2: 46%

Compared with Other (v100 Solutions)

*All validation were conducted internally (for reference only)
Cooling Alternative with Supermicro

Pure Air cooling vs. D2C liquid cooling with InRow RC

Supermicro SYS-4029GP-TVRT
Enabled Water Cooling Support

Lowest Total Power Consumption Solution
with Lowest IT Equip Power * PUE

Project Driven Solutions

Available Cooling Solutions

< PUE Driven Solution >


*Test Validated in LAB for reference only
Power Saving Differences in Liquid Cooling Environment

9x of Supermicro SYS-4029GP-TVRT

Pure Air cooling VS. D2C liquid cooling with InRow RC

Power Savings: 32% Offer Upto

Total Power Consumption / Rack:

Pure Air cooling: 46,980W

Δ = 14,920W

D2C liquid cooling with InRow RC: 32,060W

~14.7% PUE Improvement

*Data is for reference only… based on internal calculation and simulation facts

(CPU: Intel SKL-SP 6154x2 200W TDP & 768GB DDR4)
Cost Saving Over 5 Years with Water Liquid Solution

<table>
<thead>
<tr>
<th>Ranking</th>
<th>Country</th>
<th>Price (USD/kW.h)</th>
<th>Ranking</th>
<th>Country</th>
<th>Price (USD/kW.h)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Norway</td>
<td>0.042410</td>
<td>17</td>
<td>Denmark</td>
<td>0.098239</td>
</tr>
<tr>
<td>2</td>
<td>Sweden</td>
<td>0.060239</td>
<td>18</td>
<td>Greece</td>
<td>0.099310</td>
</tr>
<tr>
<td>3</td>
<td>USA</td>
<td>0.067510</td>
<td>19</td>
<td>New Zealand</td>
<td>0.100080</td>
</tr>
<tr>
<td>4</td>
<td>Luxembourg</td>
<td>0.068980</td>
<td>20</td>
<td>Turkey</td>
<td>0.105641</td>
</tr>
<tr>
<td>5</td>
<td>Mexico</td>
<td>0.070468</td>
<td>21</td>
<td>Austria</td>
<td>0.105941</td>
</tr>
<tr>
<td>6</td>
<td>Finland</td>
<td>0.073099</td>
<td>22</td>
<td>France</td>
<td>0.107569</td>
</tr>
<tr>
<td>7</td>
<td>Taiwan</td>
<td>0.075781</td>
<td>23</td>
<td>Belgium</td>
<td>0.112321</td>
</tr>
<tr>
<td>8</td>
<td>Poland</td>
<td>0.079677</td>
<td>24</td>
<td>Ireland</td>
<td>0.118640</td>
</tr>
<tr>
<td>9</td>
<td>Netherlands</td>
<td>0.083589</td>
<td>25</td>
<td>Chile</td>
<td>0.124630</td>
</tr>
<tr>
<td>10</td>
<td>Malaysia</td>
<td>0.087898</td>
<td>26</td>
<td>Philippines</td>
<td>0.124800</td>
</tr>
<tr>
<td>11</td>
<td>Hungary</td>
<td>0.088790</td>
<td>27</td>
<td>UK</td>
<td>0.125240</td>
</tr>
<tr>
<td>12</td>
<td>Czech Republic</td>
<td>0.089390</td>
<td>28</td>
<td>Portugal</td>
<td>0.125320</td>
</tr>
<tr>
<td>13</td>
<td>Israel</td>
<td>0.091020</td>
<td>29</td>
<td>Switzerland</td>
<td>0.133511</td>
</tr>
<tr>
<td>14</td>
<td>Korea</td>
<td>0.092552</td>
<td>30</td>
<td>Germany</td>
<td>0.141231</td>
</tr>
<tr>
<td>15</td>
<td>Singapore</td>
<td>0.094096</td>
<td>31</td>
<td>Japan</td>
<td>0.157890</td>
</tr>
<tr>
<td>16</td>
<td>China</td>
<td>0.096918</td>
<td>32</td>
<td>Italy</td>
<td>0.185259</td>
</tr>
</tbody>
</table>

D2C Liquid Cooling Money Saving /5 years

Supermicro Liquid cooling GPU solution Power Saving /Rack : ~14.92kW

Electricity cost difference in 5 years:

14.92kW x 24(hours) x 365(days) x 5(years) x 0.094096(USD/kW.h) = 61,491 (USD)/ Rack

Source: 2016 National Industrial Electricity Pricing Table

*Data is for reference only, 9x TVRT per RACK

© Supermicro 2018
Massive Electricity Cost Saving

The More You Scale,

The More E. Cost Saving
The More Power Saving w/ Lowest PUE

USD$ 49,523 per RACK

$$$$$$$$$$$$$$
\times N \text{ RACKS}
Optimized Low Latency GPGPU AIC Solution

High Density AIC Support in a Single System

TRT2 Series
Single Root Complex Design

Re-innovate Generic PCI-E Root Design

Provide the most optimized GPU Solutions
Dramatically improves GPU peer-to-peer communication performance

10x PCI-E 3.0 x16 (Double Width)
route under a single root complex

Optimized Parallel Thermal Design
With no GPU Pre-Heating Solution

Single Root Has 21% better Throughput, 60% better Latency over Dual Root Systems
Largest Supplier of Servers in The World

Expands its Silicon Valley Corporate Headquarters...

“...DigiTimes also reports that Supermicro has narrowed the server shipments gap with #2 manufacturer HPE to less than 1%, and has also closed on #1 supplier Dell. We are nearly 3x larger than the next largest manufacturer, Huawei, and are the fastest growing server manufacturer in the top 5. We are growing much faster than the overall market, while the market share of the very largest and the very smallest suppliers continues to erode...”

Global Server Shipment Share by Top-10 Players
Source: Digitimes Research, February 2018

ADOPTION RATE INCREASE
– Top 3 Players –
“PFN plans to enhance **MN-1** by adding **512 NVIDIA Tesla V100 32GB GPUs** and have them up and running by July, with the added GPUs having a theoretical peak performance of about 56 PetaFLOPS\(^1\), a massive 56,000 trillion floating-point operations per second, based on a mixed precision floating-point operation\(^2\) used in deep learning. This means the expansion alone will contribute to a roughly threefold increase from the current peak.”
GPU Cloud Deployment

The CSP deployed Supermicro 1U 4x GPGPU Solutions for fulfilling multiple applications in Big Data, CAE, Life Science, and M&E.

This platform allows CSP to offer various type of Accelerated cards in cloud environments…

Easily to scale with a low start-up investment per requested demands…
TVTR Series
Powered Social Media Backend

For Video/ Image and Speech Recognition Applications

Adopted Supermicro SYS-4029GP-TVRT for keeping the DATA processing demands, and offer NVLINK Max Bandwidth of 300GB/s benefit…while keeping high performance when scale-out….

Project win with FCST of 500+ for phase 1 deployment…
Deployed a total of **7x RACK HPC Compute Nodes**

Among these 7x Racks, we have successfully deployed a total of **2.5x RACK**
with **SYS-1029GQ-TXR**
(equipped NVIDIA Tesla P100)

**Deployment in East Asia**
Shipped
3234 GPU Nodes

Enhance customers’ compute and AI Infrastructures. Enriched the high performance/ density per RU designs for Search Engines, Map, Intelligent Analytic, and etc…

For East Asia Cloud Service Provider (CSP)
Widest Optimized Solutions

<table>
<thead>
<tr>
<th>Ratio: GPU:CPU</th>
<th>Tower</th>
<th>Rack</th>
<th>Deep Learning</th>
</tr>
</thead>
<tbody>
<tr>
<td>4:2 (4U)</td>
<td>4029GP-TRT 8:2 (4U)</td>
<td>1029GQ-TRT 4:2 (1U)</td>
<td>4029GP-TVRT 8:2 (4U)</td>
</tr>
<tr>
<td>4:2 (4U)</td>
<td>7049GP-TRT</td>
<td>4029GP-TVRT 8:2 (4U)</td>
<td></td>
</tr>
<tr>
<td>6:2 (2U)</td>
<td>2029GP-TR 6:2 (2U)</td>
<td>1029GQ-TRT 4:2 (1U)</td>
<td>1029GQ-TVRT 4:2 (1U)</td>
</tr>
<tr>
<td>3:2 (1U)</td>
<td>1029GP-TR 3:2 (1U)</td>
<td>1019GP-TT 2:1 (1U)</td>
<td></td>
</tr>
<tr>
<td>2:1 (1U)</td>
<td>1019GP-TT 2:1 (1U)</td>
<td>5019GP-TT 2:1 (1U)</td>
<td>4029GP-TRT2 8:2 (4U)</td>
</tr>
</tbody>
</table>
Please visit our Supermicro booth, and learn more about our Solutions...
Thank You

https://www.supermicro.com/products/nfo/gpu.cfm

For more information, visit www.supermicro.com or contact marketing@supermicro.com
Disclaimer
Super Micro Computer, Inc. may make changes to specifications and product descriptions at any time, without notice. The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions and typographical errors. Any performance tests and ratings are measured using systems that reflect the approximate performance of Super Micro Computer, Inc. products as measured by those tests. Any differences in software or hardware configuration may affect actual performance, and Super Micro Computer, Inc. does not control the design or implementation of third party benchmarks or websites referenced in this document. The information contained herein is subject to change and may be rendered inaccurate for many reasons, including but not limited to any changes in product and/or roadmap, component and hardware revision changes, new model and/or product releases, software changes, firmware changes, or the like. Super Micro Computer, Inc. assumes no obligation to update or otherwise correct or revise this information.

SUPER MICRO COMPUTER, INC. MAKES NO REPRESENTATIONS OR WARRANTIES WITH RESPECT TO THE CONTENTS HEREOF AND ASSUMES NO RESPONSIBILITY FOR ANY INACCURACIES, ERRORS OR OMISSIONS THAT MAY APPEAR IN THIS INFORMATION.

SUPER MICRO COMPUTER, INC. SPECIFICALLY DISCLAIMS ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR ANY PARTICULAR PURPOSE. IN NO EVENT WILL SUPER MICRO COMPUTER, INC. BE LIABLE TO ANY PERSON FOR ANY DIRECT, INDIRECT, SPECIAL OR OTHER CONSEQUENTIAL DAMAGES ARISING FROM THE USE OF ANY INFORMATION CONTAINED HEREIN, EVEN IF SUPER MICRO COMPUTER, Inc. IS EXPRESSLY ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

ATTRIBUTION
© 2018 Super Micro Computer, Inc. All rights reserved.