Open. Together.
Hardware Specification Contribution:
- Glacier Point V2
- M.2 Accelerator
- Dual M.2 Accelerator

Hao Shen
hao1shen@fb.com
Facebook
Hardware Acceleration

Machine Learning - Inference

Video Transcoding
How does FB use Machine Learning?

- Search
- Translation
- Recognition
- News Feed
- Ads

Open. Together.
Facebook AI Hardware

Today

- **Bryce Canyon**
  - High density storage
  - Storage-heavy tasks

- **Big Basin**
  - 8 NVIDIA P100s/V100s
  - Very compute-heavy tasks

- **Tioga Pass**
  - Dual socket, many cores CPU
  - Compute & mem-heavy tasks

- **Twin Lakes**
  - Single socket, fewer cores CPU
  - Compute & low latency tasks

---

**Deployment Platform**

Artificial Intelligence Framework

Backend: Compiler/Optimizer/Library

**Hardware Infrastructure**

---

**Open. Together.**
Facebook Accelerator Module – M.2 and Dual M.2 form factor

Standard M.2 Form Factor
- ASIC
- DRAM
- 12W TDP
- PCIe x4
- UART/JTAG

Dual M.2 Form Factor
- ASIC
- DRAM
- 20W TDP
- PCIe x8
- UART/JTAG
M.2 Accelerator Module - PCBA

- 22x110mm
- 0.8mm thick
- Up to 12 layers
- 2.5mm height on top
- 1.5mm height on bottom
M.2 Accelerator Module – Power, Interface

- 10W TDP (module)
- LPDDR4x 64bit x2
- PCIe x4
- SMBus
- UART
- JTAG
- USB*
Dual M.2 Accelerator Module - PCBA

- 46x110mm
- 0.8mm thick
- Up to 12 layers
- 2.5mm height on top
- 1.5mm height on bottom
Dual M.2 Accelerator Module - Connector
Dual M.2 Accelerator Module – Power, Interface

- 20W TDP (module)
- LPDDR4x 64bit x4
- PCIe x8
- SMBus
- UART
- JTAG
- USB*

Main Slot (Slot A)

<table>
<thead>
<tr>
<th>Signal</th>
<th>Pin 1</th>
<th>Pin 2</th>
<th>Pin 3</th>
<th>Pin 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.3V</td>
<td>A2</td>
<td>A3</td>
<td>A4</td>
<td>A5</td>
</tr>
<tr>
<td>3.3V</td>
<td>A6</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
</tr>
<tr>
<td>PHARES</td>
<td>A7</td>
<td>PEIns</td>
<td>PEIns</td>
<td>PEIns</td>
</tr>
<tr>
<td>LDO 4.0V</td>
<td>A8</td>
<td>A9</td>
<td>A10</td>
<td>A11</td>
</tr>
<tr>
<td>3.3V</td>
<td>A12</td>
<td>A13</td>
<td>A14</td>
<td>A15</td>
</tr>
<tr>
<td>3.3V</td>
<td>A16</td>
<td>A17</td>
<td>A18</td>
<td>A19</td>
</tr>
<tr>
<td>THST</td>
<td>A20</td>
<td>A21</td>
<td>A22</td>
<td>A23</td>
</tr>
<tr>
<td>TDO</td>
<td>A24</td>
<td>A25</td>
<td>PEIns</td>
<td>PEIns</td>
</tr>
<tr>
<td>TDI</td>
<td>A26</td>
<td>A27</td>
<td>PEIns</td>
<td>PEIns</td>
</tr>
<tr>
<td>TDO</td>
<td>A28</td>
<td>A29</td>
<td>A30</td>
<td>A31</td>
</tr>
<tr>
<td>SMB (SMB1.9)</td>
<td>A32</td>
<td>A33</td>
<td>A34</td>
<td>A35</td>
</tr>
<tr>
<td>SMB DATA (SMB1.8)</td>
<td>A36</td>
<td>A37</td>
<td>A38</td>
<td>A39</td>
</tr>
<tr>
<td>ALERT (SMB1.7)</td>
<td>A40</td>
<td>A41</td>
<td>A42</td>
<td>A43</td>
</tr>
<tr>
<td>Reserved</td>
<td>A44</td>
<td>A45</td>
<td>A46</td>
<td>A47</td>
</tr>
<tr>
<td>Reserved</td>
<td>A48</td>
<td>A49</td>
<td>A50</td>
<td>A51</td>
</tr>
<tr>
<td>Reserved</td>
<td>A52</td>
<td>A53</td>
<td>A54</td>
<td>A55</td>
</tr>
<tr>
<td>Reserved</td>
<td>A56</td>
<td>A57</td>
<td>A58</td>
<td>A59</td>
</tr>
<tr>
<td>Reserved</td>
<td>A60</td>
<td>A61</td>
<td>A62</td>
<td>A63</td>
</tr>
<tr>
<td>Reserved</td>
<td>A64</td>
<td>A65</td>
<td>A66</td>
<td>A67</td>
</tr>
<tr>
<td>Reserved</td>
<td>A68</td>
<td>A69</td>
<td>A70</td>
<td>A71</td>
</tr>
<tr>
<td>Reserved</td>
<td>A72</td>
<td>A73</td>
<td>A74</td>
<td>A75</td>
</tr>
</tbody>
</table>

2nd Slot (Slot B)

<table>
<thead>
<tr>
<th>Signal</th>
<th>Pin 1</th>
<th>Pin 2</th>
<th>Pin 3</th>
<th>Pin 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.3V</td>
<td>B2</td>
<td>B3</td>
<td>B4</td>
<td>B5</td>
</tr>
<tr>
<td>3.3V</td>
<td>B6</td>
<td>B7</td>
<td>B8</td>
<td>B9</td>
</tr>
<tr>
<td>PHARES</td>
<td>B10</td>
<td>B11</td>
<td>B12</td>
<td>B13</td>
</tr>
<tr>
<td>LDO 4.0V</td>
<td>B14</td>
<td>B15</td>
<td>B16</td>
<td>B17</td>
</tr>
<tr>
<td>3.3V</td>
<td>B18</td>
<td>B19</td>
<td>B20</td>
<td>B21</td>
</tr>
<tr>
<td>SMB (SMB1.9)</td>
<td>B22</td>
<td>B23</td>
<td>B24</td>
<td>B25</td>
</tr>
<tr>
<td>SMB DATA (SMB1.8)</td>
<td>B26</td>
<td>B27</td>
<td>B28</td>
<td>B29</td>
</tr>
<tr>
<td>ALERT (SMB1.7)</td>
<td>B30</td>
<td>B31</td>
<td>B32</td>
<td>B33</td>
</tr>
<tr>
<td>Reserved</td>
<td>B34</td>
<td>B35</td>
<td>B36</td>
<td>B37</td>
</tr>
<tr>
<td>Reserved</td>
<td>B38</td>
<td>B39</td>
<td>B40</td>
<td>B41</td>
</tr>
<tr>
<td>Reserved</td>
<td>B42</td>
<td>B43</td>
<td>B44</td>
<td>B45</td>
</tr>
<tr>
<td>Reserved</td>
<td>B46</td>
<td>B47</td>
<td>B48</td>
<td>B49</td>
</tr>
<tr>
<td>Reserved</td>
<td>B50</td>
<td>B51</td>
<td>B52</td>
<td>B53</td>
</tr>
<tr>
<td>Reserved</td>
<td>B54</td>
<td>B55</td>
<td>B56</td>
<td>B57</td>
</tr>
<tr>
<td>Reserved</td>
<td>B58</td>
<td>B59</td>
<td>B60</td>
<td>B61</td>
</tr>
<tr>
<td>Reserved</td>
<td>B62</td>
<td>B63</td>
<td>B64</td>
<td>B65</td>
</tr>
<tr>
<td>Reserved</td>
<td>B66</td>
<td>B67</td>
<td>B68</td>
<td>B69</td>
</tr>
<tr>
<td>Reserved</td>
<td>B70</td>
<td>B71</td>
<td>B72</td>
<td>B73</td>
</tr>
<tr>
<td>Reserved</td>
<td>B74</td>
<td>B75</td>
<td>B76</td>
<td>B77</td>
</tr>
</tbody>
</table>

* USB includes USB 3.1 and USB 3.2.
M.2/Dual M.2 Accelerator Module Specification Contents

- Power transient spec
- PCIe details
- Reset Architecture
- FRU information
- Reliability
- Compliance
Facebook Accelerator System

Module
- Single M.2
- Dual M.2

Glacier Point v2 Card
- Up to 12x single/
  6x dual M.2

Yosemite V2
- 2x TwinLakes with 2x Glacier
  Point V2 Cards
YV2 Platform Block Diagram

CPU

System Memory

Networking

PCle Switch

Accelerator
Accelerator
Accelerator
Accelerator

GPV2

PCle Switch

Accelerator
Accelerator
Accelerator
Accelerator

System Memory

DDR4

Supernode

Open. Together.
GPv2 PCIe Block Diagram
GPv2 I2C Block Diagram