
OPEN
Compute Project
OCP U.S. SUMMIT 2017
Santa Clara, CA

OCP Mezzanine NIC 2.0 & Beyond

@OCP Server Workgroup track

Jia Ning, Hardware Engineer

Facebook

3/9/2017

OPEN HARDWARE. OPEN SOFTWARE. OPEN FUTURE.

OCP Mezzanine NIC 3.0 Discussion

Rev 03

2/16/2017

Jia Ning, Hardware Engineer, Facebook

Background

OCP Mezzanine Cards

All accepted by the IC

Specification	Version	Submit Date	Contributor	License	Notes
OCP Mezzanine card v2.0 OCP_Mezz_2.0_rev1.00_20151215b_pub_release.pdf(2.2MB) OCP_Mezz_2.0_rev1.00_20151215b_pub_release_3D_package.zip (88MB) Mechanical 20151023_P1-P9_K1-K5 zip file (57MB)	V2.0-1.0	Dec 15, 2015	Facebook	OWFa 1.0	Added support for x16 (quad x4), NCSI, dual QSFP+, & Quad SDP+ Accepted by OCP IC 2/24/2016
OCP Mezzanine card v0.5, original standard Mezzanine Card (rev 0.5)	V0.5	Oct 8, 2012	Facebook	OWFa 1.0	Defacto standard for the original network mezzanine with a x8 PCIe Gen3 interface

OCP Mezz v0.5 defined ~4-5 years ago:

- 10G Ethernet
- 2x SFP
- X8 PCIe Gen3
- I2C sideband

OCP Mezz v2.0 defined ~1-2 years ago:

- 10/25/40/50/100G Ethernet
- Up to 4x SFP28, 2x QSFP28, 4x RJ45
- X16 PCIe Gen3
- NCSI Sideband

Status

- In general, the community is seeing healthy adoption on both the NIC side and system side
- Host side connection has path to Gen4 16Gbps

http://www.fci.com/en/products/board-to-board-wire-to-board/board-to-board/08mm-board-to-board-signal/bergstak-plus-08mm-pcie-4-mezzanine.html

- Receiving many inquiries for implementation detail
- Receiving feedback for “pain points”

Examples of adopters of OCP Mezz NIC form factor:

Broadcom
https://www.broadcom.com/products/ethernet-connectivity/network-adapters/ocm14102-nx-ocp

Chelsio
http://www.chelsio.com/nic/unified-wire-adapters/t580-ocp-so/

Intel
http://www.intel.com/content/www/us/en/ethernet-products/converged-network-adapters/server-adapter-x520-da1-da2-for-ocp-brief.html

Mellanox
http://www.mellanox.com/ocp/index.php

Qlogic
http://www.qlogic.com/Resources/Documents/DataSheets/Adapters/Datasheet_QOE2562_Adapters.pdf

Quanta
https://www.qct.io/product/index/Server/Server-Accessory/OCP-Network-Mezzanine-For-Server

Silicom
http://www.silicom-usa.com/cats/server-adapters/ocp-mezzanine-adapters/

Wiwynn
http://www.wiwynn.com/english/product/type/details/59?ptype=37

Zaius (Rackspace/Google)
http://files.opencompute.org/oc/public.php?service=files&t=d99c1c5aac68df38e09856f5c6e96a13&download

“Pain Points” and Problem Statement

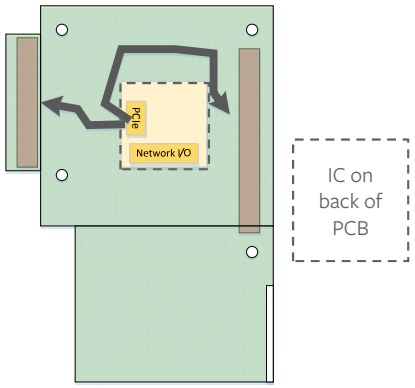
- Gates emerging use cases
- Blocks further expansion of adoption
- Understand the problem and solve by making changes to Mezz 2.0 specification

- Board space is not enough for:
  - Larger package IC
  - Multi-IC solution (NIC + FPGA/processor)
  - NIC with external DRAM
  - Higher I/O bandwidth (more connector BW/count)
  - Potential x32 PCIe
  - Lack of length/width tiers like PCIe LP/FH-HL/FH

- Mechanical profile is not enough for:
  - 10~20+W (100G NIC or other type of IC) vs. 3-5W(10G NIC)
  - High ambient use cases(rear I/O, high ambient data center)
  - Some optical module use cases

More “Pain Points”

- Connector placement is not routing friendly
  - Connector location at opposite sides of card makes routing challenging
  - PCIe routing and DRAM routing is crossing
- Specification has usability challenge
  - Concern about connector compatibility risks; hard to navigate through connector A,B,C and type 1,2,3,4
  - Specification is incremental, and need some background of previous specifications to understand
- Lack of common EMI plate to allow chassis I/O to take different Mezz NIC as FRU



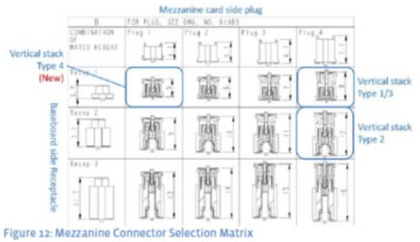
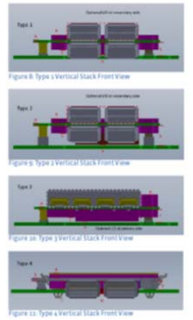


Figure 12: Mezzanine Connector Selection Matrix



						/Secondary side optional	
TYPE 2	1.57mm	2.9mm /2.0mm	4.5mm /4mm	11.5mm	12mm	Primary side /Secondary side optional	Primary side
TYPE 3	1.57mm	7.5mm	4.5mm /4mm	7.5mm	8mm	Primary side optional /Secondary side	Primary side/ secondary side
Type 4	1.57mm	2.9mm /2mm	4.5mm /4mm	7.5mm	5mm	Primary side /Secondary side optional	Primary side

Why not just use PCIe?

- We ask ourselves this question during survey to calibrate whether we are seeking problems for a solution.
- Limitations of PCIe CEM form factor exist:
  - **Not able to use NC-SI sideband** – *valuable for shared NIC in all power states*
  - **Not compatible with Multi-host NIC requirements** - *such as 4x clocks*
  - **Power domain difference and standby power is limited** – *NIC tends to be active/partially active during S5*
  - **Compact size of Mezz is preferred** - often provide 1x extra slot for system configuration
- An OCP Mezz NIC spec with the above limitation addressed has value for NIC/system/CSP

Mezz 3.0 General Approach

- Understand the problem
  - Collect feedback from Internal, NIC vendors, system vendors, CSP
  - Talk to NIC and system vendors to understand use cases
  - Target to unblock new use cases and thermal challenge, with migration challenge considered

- Find and implement a solution
  - Work on NIC form factor change proposal under OCP Mezz NIC subgroup
  - Form consensus in work group and finalize specification change and migration plan
  - Leave enough time to impact the planning for next generation NICs cards and systems

Mezz 3.0 Migration Community Feedback

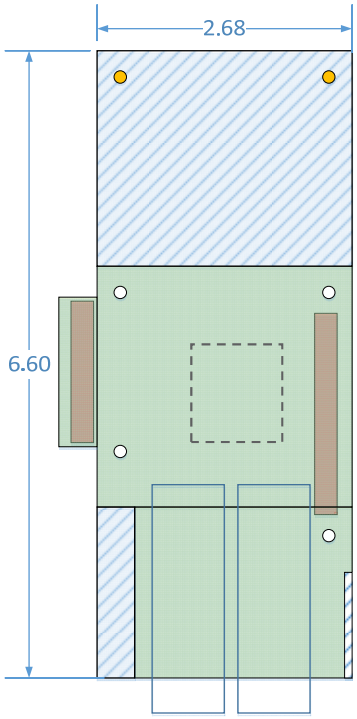
Impact to NIC	Priority
Use case – higher Thermal	High
Use case - Larger IC package	Medium
Use case - Multi-chip card	Medium
Use case - IC w/ DRAM	Medium
Stable form factor considering needs in the next 5 years	High
Impact to system	
Feasibility for New System mechanical and board design	High
Feasibility to migrate existing system mechanical design to support Mezz 3.0	Medium
Feasibility to migrate exist board design to support Mezz 3.0	Low
Impact to ecosystem migration	
Existing baseboard to support Mezz 3.0	Low
Plan to allow system and card to migrate to Mezz 3.0	High

Summary of Options

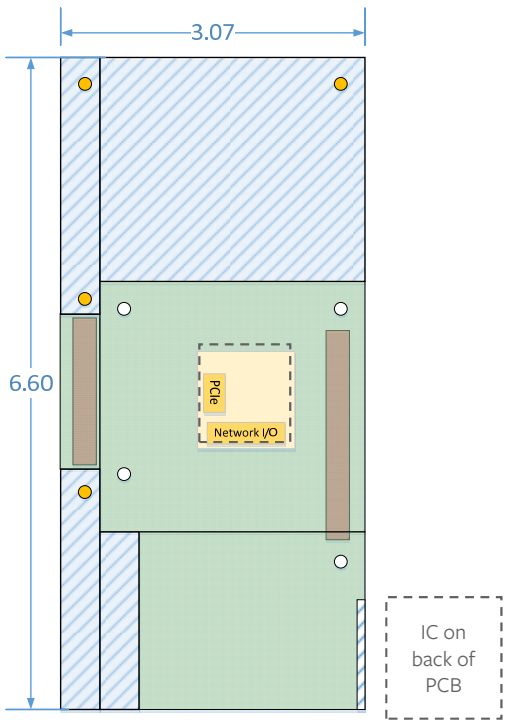
	Description of change made to Mezz 2.0	Status	Feedback
1	Add 16mm stacking height option	Close	NIC placement challenge has no improvement
2	Extend width	Close	NIC placement challenge has no improvement
3	Extend length	Open	NIC Placement challenge is partially addressed NIC PCIe routing challenge still exist
4	Move connector B to right edge	Close	NIC PCIe routing challenge gets worse
5	Flip ASIC to top	Close	Force tradeoff between system configuration flexibility and thermal
6	Move connector A to the left edge and keep same Y-location as Mezz 2.0	Open	NIC PCIe layout has crossing Possible backward compatibility for x8
7	Move connector A to the left edge with smooth PCIe lane sequence	Open	Best option for long term Lack of backward compatibility Most challenging for migration
8	Based on 7 and turn connector B by 180°	Open	NIC PCIe layout has crossing Possible backward compatibility for x16 with dual layout

Enumeration of #3

Option 3a
Extend length only



Option 3b
Extend length and fill up the space



Pros:

- Added PCB length helps with some new NIC use cases
- Able to fit 4x SFP+ and 4x RJ45 (3b only)

Cons:

- More depth adds challenge to new board design to support Mezz 3.0
- More complicated to design new baseboard to have mounting holes for both Mezz2.0 and Mezz3.0 (3b only)
- Possible long PCIe trace adds risk to PCIe Gen4 SI

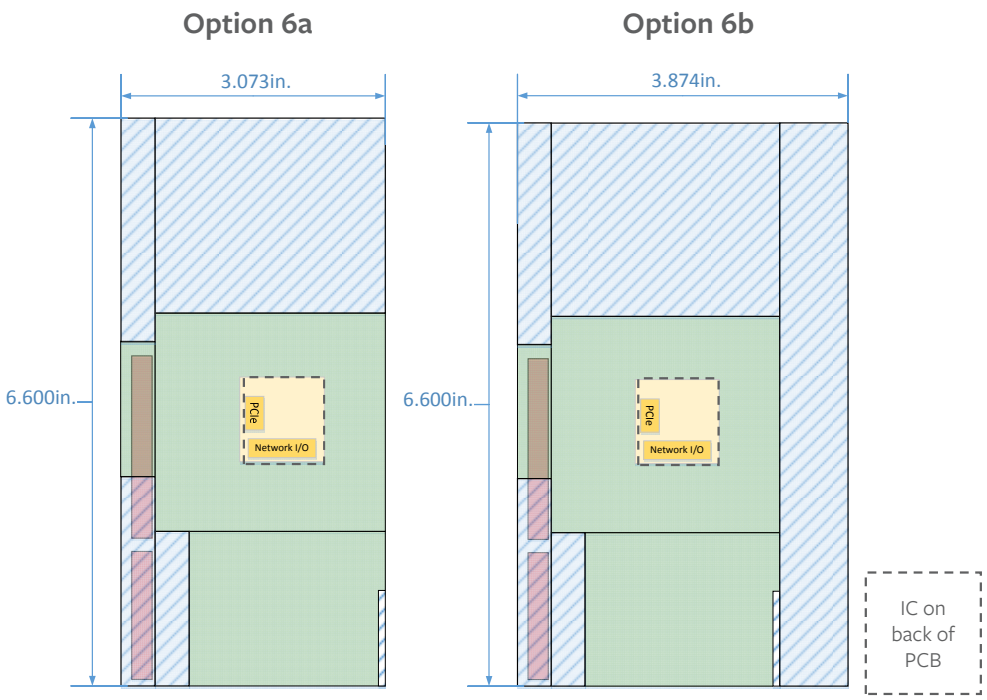
Feedback:

- #1 from a NIC vendor:
  - NIC + FPGA (up to 40x40) + 5x DRAM + 2x QSFP application is able to fit in 14 layer stack
- #2 from a NIC vendor:
  - SoC (45x45) with 10x DRAM has a PCIe breakout challenge
  - Routing of DRAM is blocked by PCIe
- #3 from a NIC vendor
  - PCIe routing direction in the way of DRAM routing for SoC + 9x DRAM application
- #4 from a CSP:
  - Need size close to FH PCIe

FH-HL 4.2" x 6.8" (3.9" x 6.6" usable) = 25.74 sq in.
3b 3.07 x 6.6 (-10%) = 18.2 sq in -> 30% less real estate

Enumeration of #6

- Move Connector A to the side of Connector B
- Make 2x width/2x length option
- Place Connect A in the same Y location as Mezz2.0



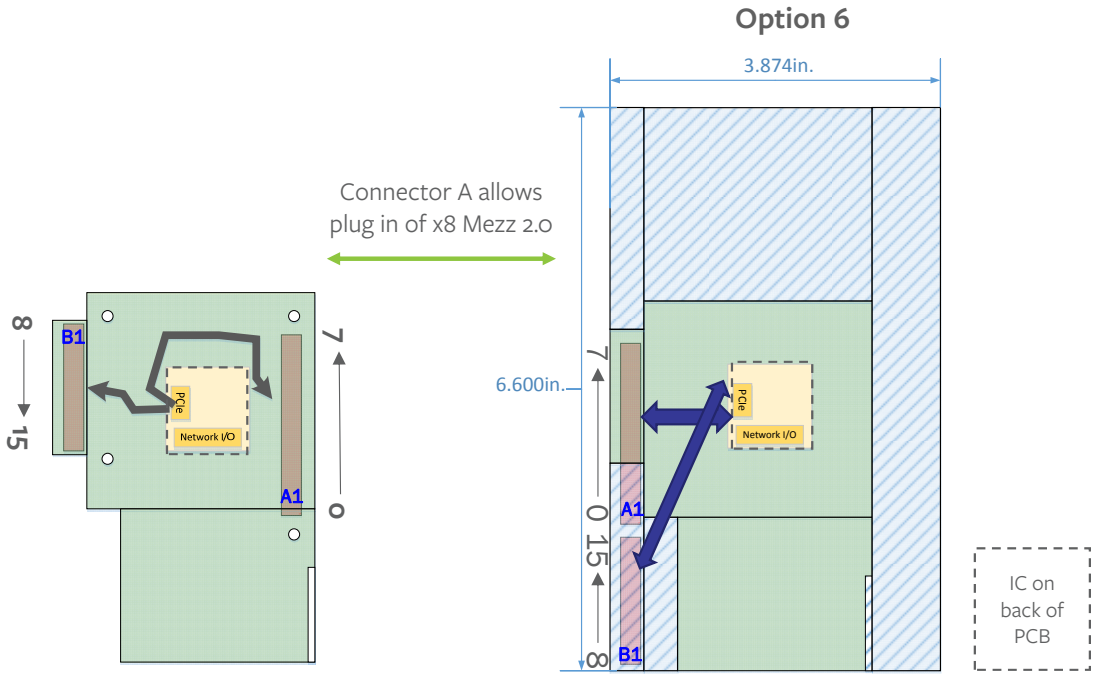
Pros:

- Helpful for thermal (heatsink zone will be enlarged)
- Helpful for placement
  - PCIe routing is short
  - DRAM placement is feasible
  - Accommodate larger package
- Potential to add one connector for x32 use cases
- Possible backward compatibly for x8 card by placing Connector A at the same “Y” location

Cons – *continued on next slide…*

Enumeration of #6

- Move Connector A to the side of Connector B
- Make 2x width/2x length option similar to PCIe CEM
- Put Connector A in the same “Y” location allows possible baseboard design to accommodate both Mezz 2.0 and Mezz3.0

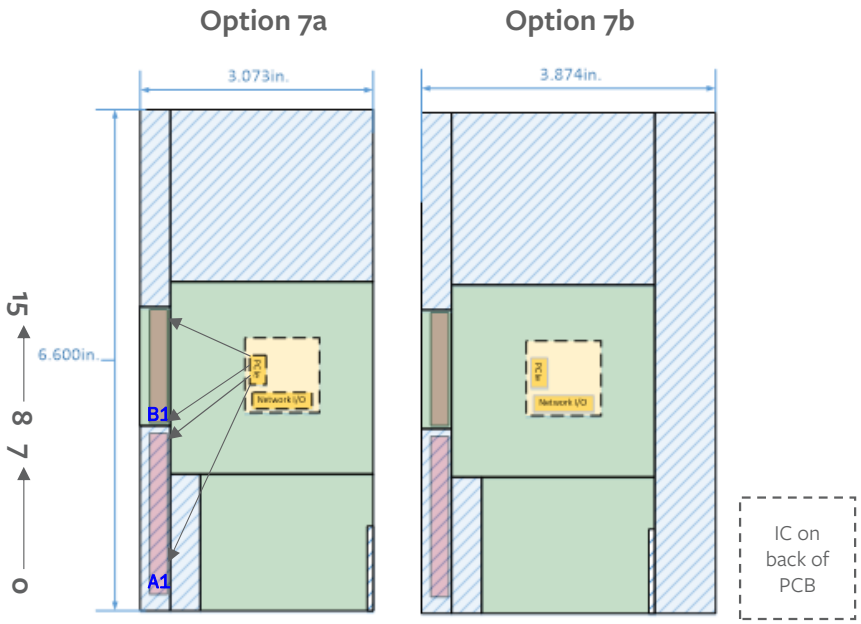


Cons:

- Possible routing challenge to be solved for Mezz 2.- mounting hole pattern at baseboard overlapping with Connector B in Mezz 3.0
- Upper and lower x8 PCIe routing are crossing
- May drive layer count depending on breakout plan and total routing layers available
  - NIC vendor input is needed
- Adds Risk to PCIe Gen4 SI

Enumeration of #7

- Move Connector A to the side of Connector B
- Make 2x width/2x length options
- Place A1 and B1 for same PCIe lane sequence as PCIe CEM gold finger



Pros:

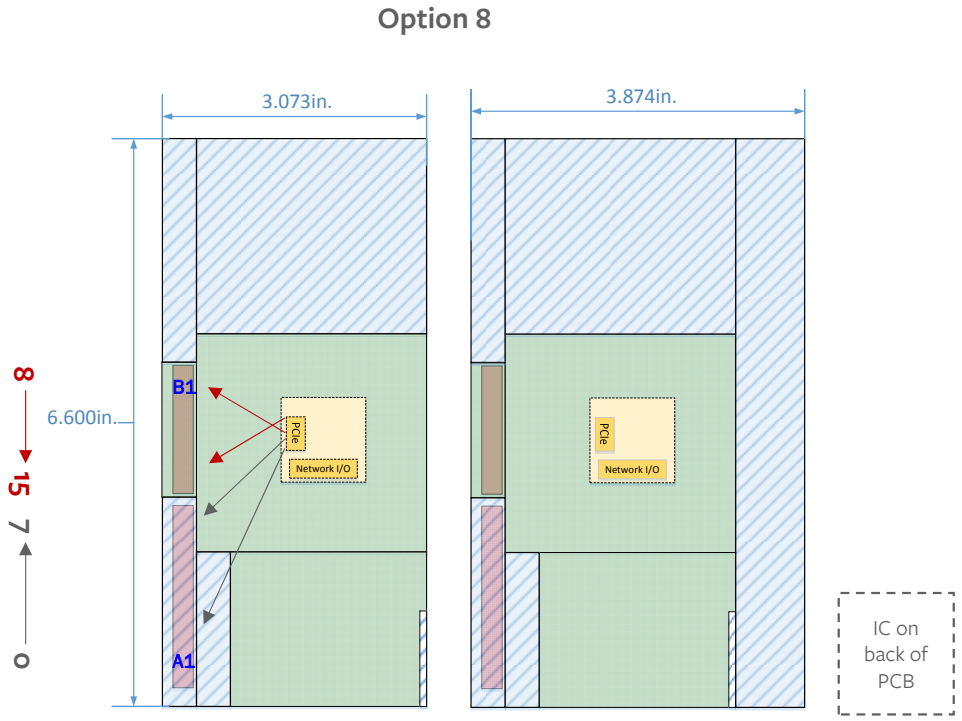
- Carries most of the mechanical benefits from option 6 thermal
- Easy modification from PCIe card for NIC vendors
- A good option for stable form factor

Cons:

- Not able to support current Mezz 2.0
  - Force system vendors to convert without path for backward compatibility and increase the friction of adoption greatly
  - Increase frication of NIC vendor's adoption due to lack of supporting systems
- Needs NIC vendor input on:
  - Is this the best option for long term NIC planning?
  - Willingness to support both Mezz 2.0 and Mezz3.0 form factor in a period to be defined, to encourage system vendors' adoption of Mezz 3.0?

Enumeration of #8

- Same as #7, except that Pin B1 location is changed to be same as Mezz 2.0



Pros:

- Share most pros from option 6 on thermal and placement wise
- Allows possible baseboard co-layout of Connector B for x16 Mezz 2.0

Cons:

- Possible complication of PCIe breakout for 8 lanes in x16 NIC
  - NIC vendor input is needed
- Adds Risk to PCIe Gen4 SI
  - NIC vendor input is needed
- Challenge with Mezz 2.0 mounting hole pattern hits blocks new Connector A's breakout

Next Steps – 2017/2/16

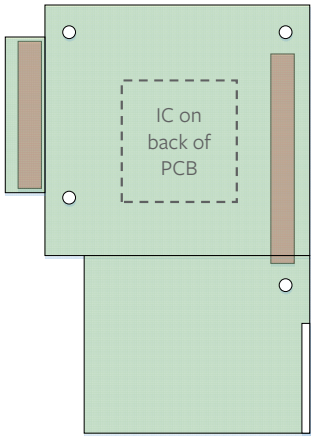
- One more round of feedback collection from System/NIC/CSPs
- Work on other feedback that has yet to be addressed by the form factor change
- Carry on activities and make progress in OCP Mezz NIC subgroup
  - Wiki: http://www.opencompute.org/wiki/Server/Mezz
  - Mailing list: http://lists.opencompute.org/mailman/listinfo/opencompute-mezz-card
  - Mezz Subgroup calls: http://opencompute.org/community/ocp-calendars
  - Workshops: TBD

Backups

Enumeration #1

Option 1

Increase Z-height from 8/12mm to 16mm



Propose to put aside

Pros:

- Most effective to help with the thermal challenge
- Lowest impact to board design and compatibility
- Can co-exist with other options

Cons:

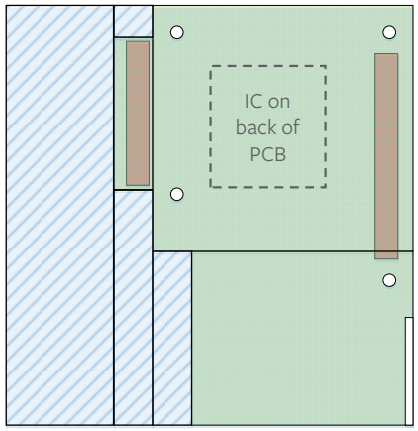
- Only increases Z-height and is not able to help with other use cases
- Higher profile occupies more space and limits the system level configuration flexibility
- 16mm has higher risk on PCIe Gen4 SI

Propose to put aside due to not addressing placement which is a major pain point

Enumeration #2

Option 2

Keep connector location and Extend width



Propose to put aside

Pros:

- Maximizes the I/O area

Cons:

- Connector B is in the middle and is not able to utilize the extended space well for new NIC use cases
- Takes space from the motherboard's onboard device's I/O

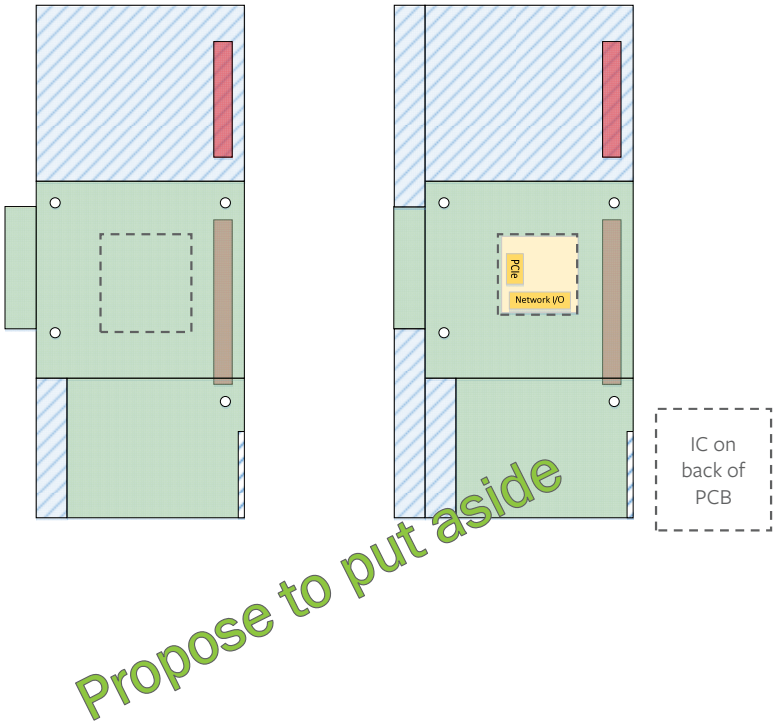
Take off this option due to lack of benefit

Enumeration #4

Move Connector B to same side as Connector A

Option 4a

Option 4b



Cons:

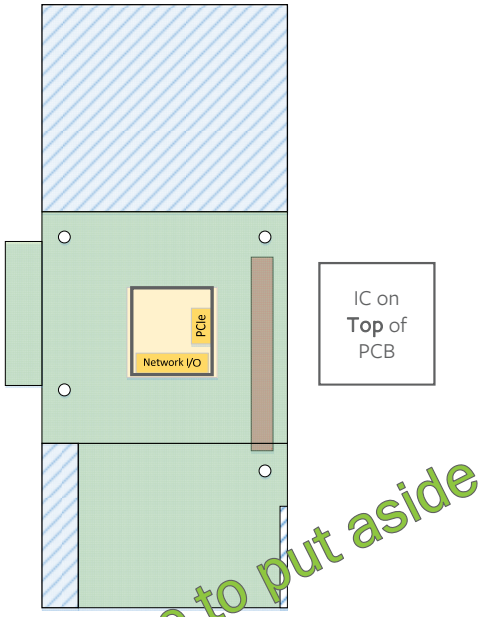
- Moving connector B location prevents the new system from being able to support current Mezz 2.0
- Placement-wise, it makes PCIe routing very unfavorable and very challenging on DRAM routing

Put aside this option due to negative impact without good enough long term benefit

Enumeration #5

Flip the ASIC to top side

Option 5



When Mezz MH = 5mm, how tall keep-out definition is acceptable?

	Option Q1: 2OU / PCIe cardx3 populated above Mezz	Option Q2: 1RU / PCIe cardx1 populated above Mezz	Option Q3: 2RU / PCIe cardx3 populated above Mezz
Platform A (2OU)	4.9mm	None	None
Platform B (2OU)	4.9mm	None	None
Platform C (2OU)	5.9mm	None	None
Platform D (2OU)	5.9mm	None	None
Platform E (2OU)	4.6mm	None	None
Platform F (1RU)	None	5.2mm	None
Platform G (2RU)	None	None	4.29mm
Platform H (1RU)	None	9.9mm	None
Platform I (1RU)	None	5.7mm	None
Platform J (2RU)	None	None	4.6mm

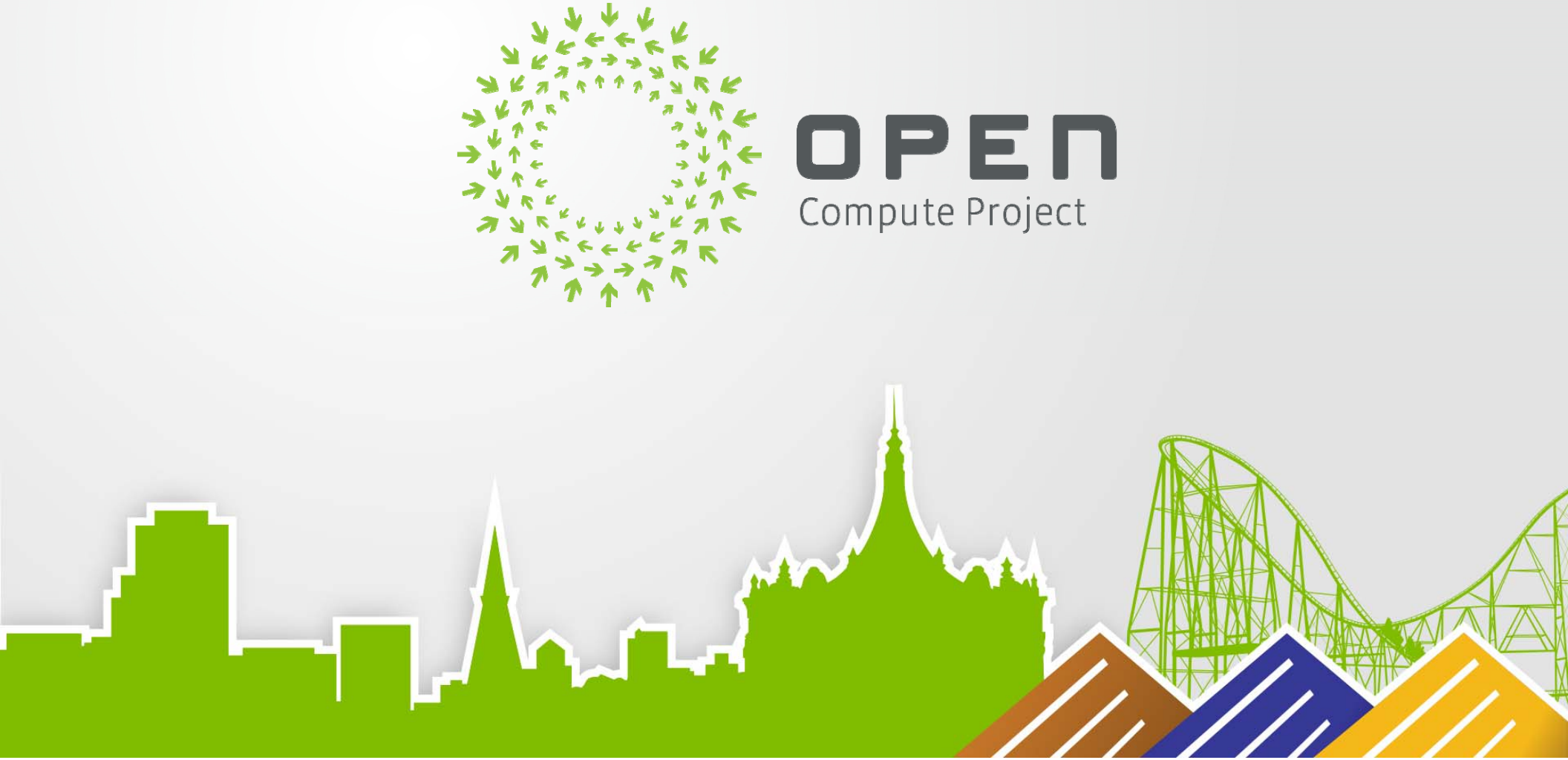
Pros:

- Solves the PCIe routing direction issue
- Retains connector A/B location for best baseboard backward compatibility

Cons:

- Mezz 3.0 heatsink height is very limited and even shorter than what Mezz2.0

Suggest to put this option aside


OPEN
Compute Project