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

AVA: NVMe M.2 in Scale-Out Storage

Dominic Cheng / Hardware Engineer / Facebook

Michael Liberte / Partner Engineer / Facebook

OPEN HARDWARE. OPEN SOFTWARE. OPEN FUTURE.

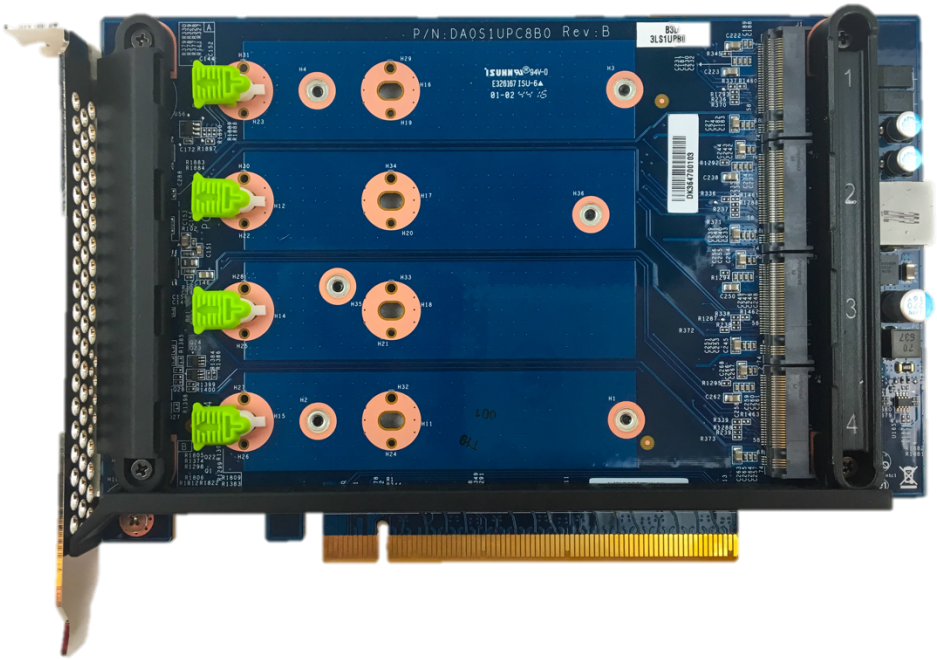
OPEN Compute Project

Agenda

- Design objectives
- Design overview
- Use cases
- Disaggregate Lab

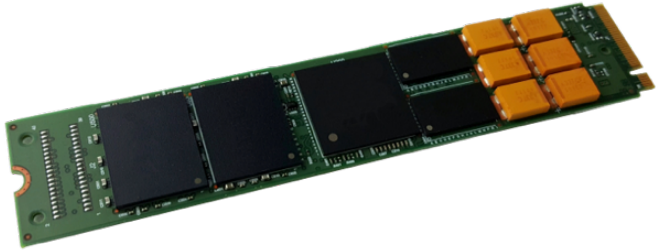
Design objectives

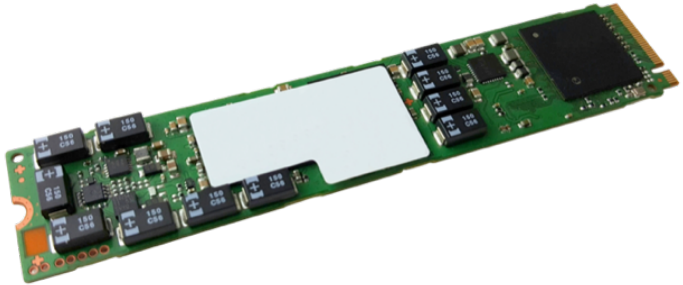
- Build a PCIe flash card with M.2s
- M.2 into PCIe FHHL form factor
- No PCIe switches or re-drivers
- Future-proof
- Serviceable
- Low cost



Why M.2?

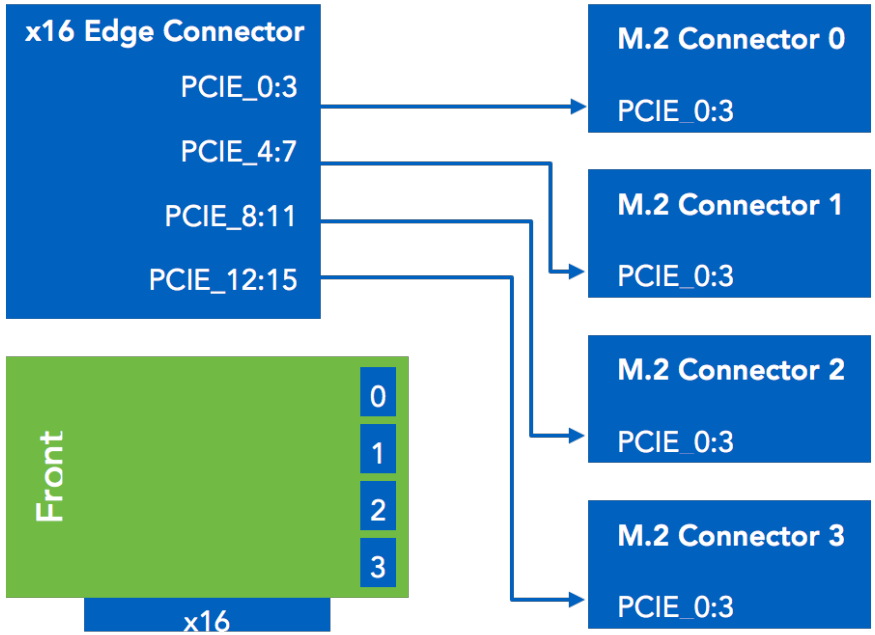
- Fits into all of our storage and compute platforms
- Commodity
- Scales





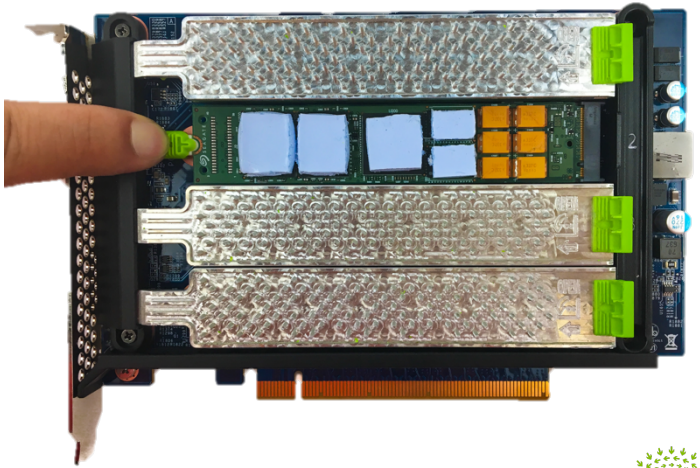
Design overview

- Direct connection from x16 edge connector to 4 M.2 connectors
- Standard Socket 3, M-key pin-out
  - x4 lanes to each



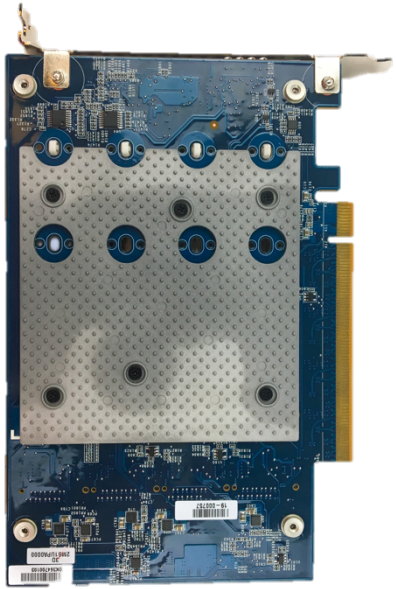
Mechanical

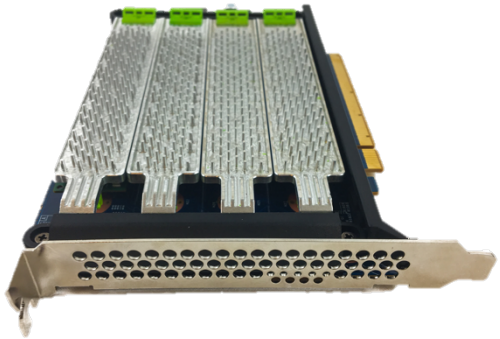
- Supports 2280 and 22110 lengths
- 5.8 mm height connectors
  - Supports TIMs
  - Up to D5 height modules
- Top and bottom heatsinks
- Tool-less service

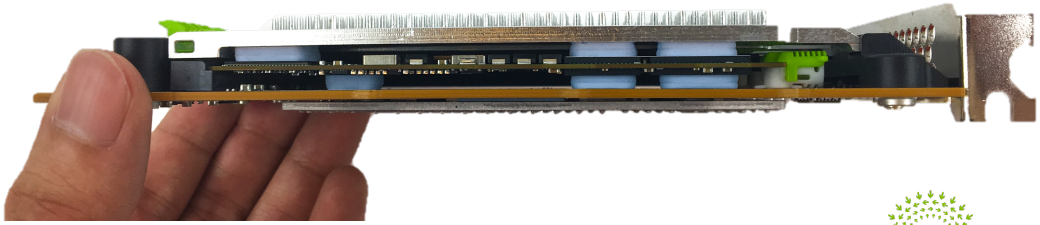


Thermal

- Top and bottom heatsinks
- TIMs for heat transfer





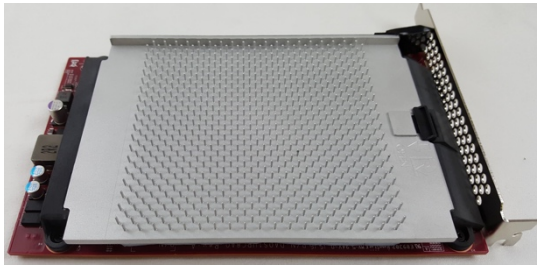


Thermal

Stress/Load	M.2 SMART Reading, 30C Inlet		
	AVA + Air Duct, 2X M.2	AVA + Heat Sink, 2X M.2	Power Per M.2
100% Sequential Writes	74 ˚C	54 ˚C	7.9 W

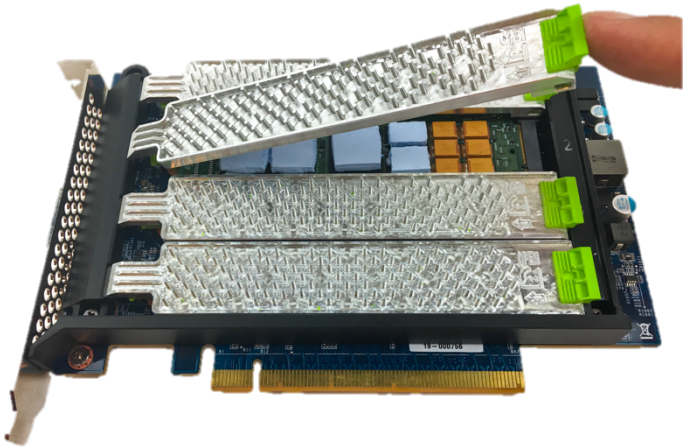
AVA + Air Duct, 2X M.2

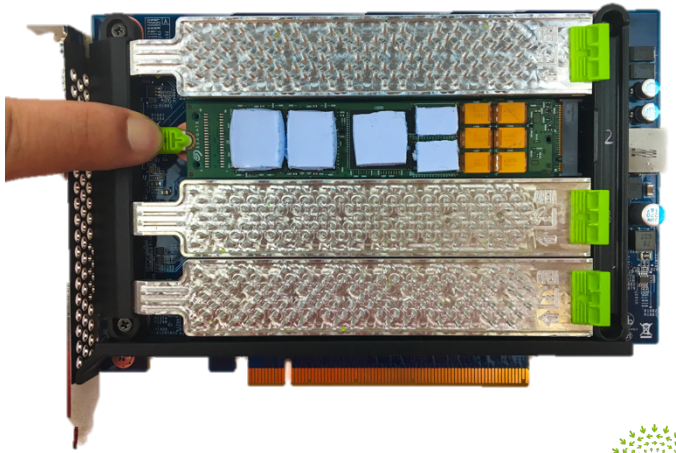
AVA + Heat Sink, 2X M.2



Serviceability

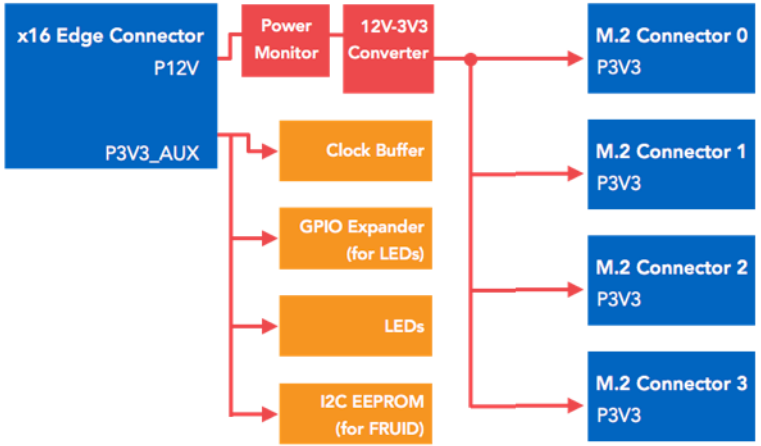
- Tool-less replacement of **individual** M.2 and top heat-sinks
- TIMs replaced with M.2





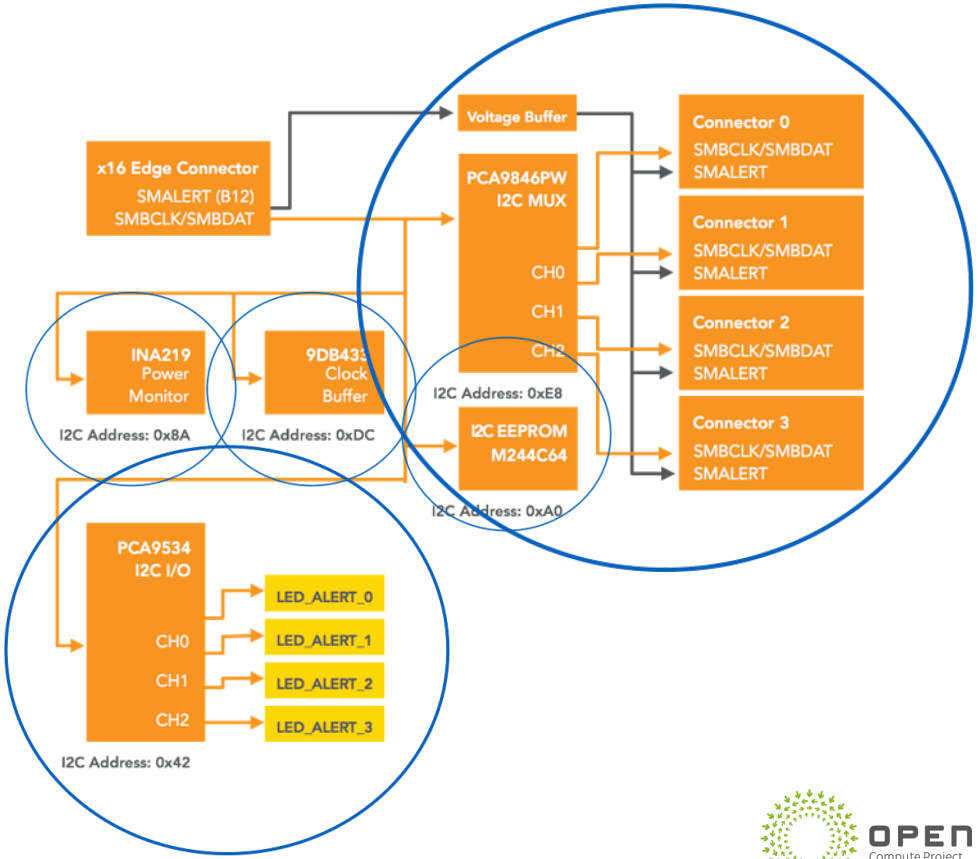
Power

- 3.3 VR supports up to 18A of total continuous current
- All I2C/SMBus devices powered from 3.3Vaux



I2C/SMBUS

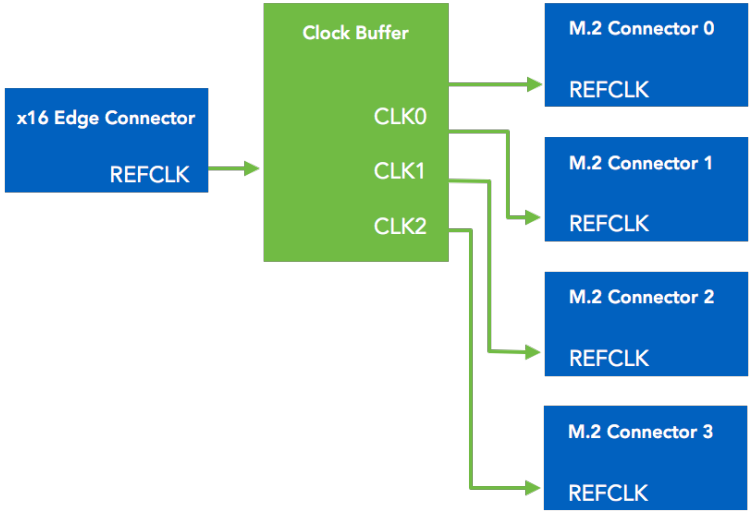
- SMBus connection to each M.2
- FRU EEPROM
- Status LED control
- Power monitor
- Clock buffer

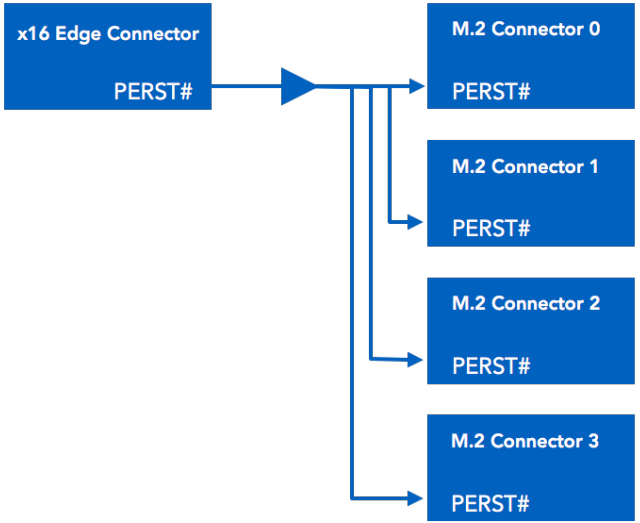


OPEN HARDWARE. OPEN SOFTWARE. OPEN FUTURE.

OPEN Compute Project

Clocks and Reset





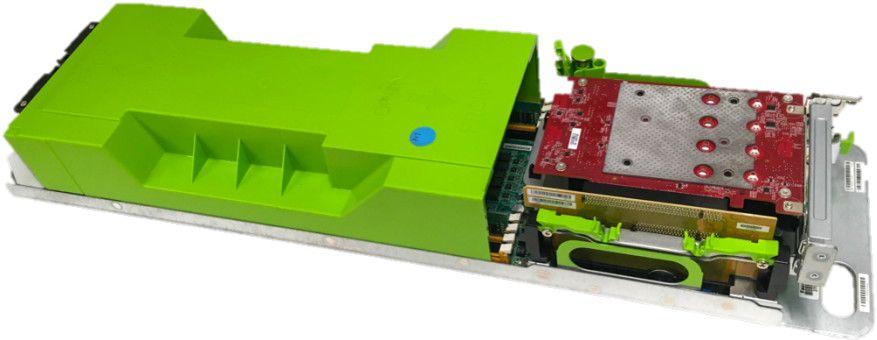
PCIe Bifurcation

- Pin B31 is pulled low on the card
- Pin B81 is connected to pin A1 on the card for presence detection
- All PRSNT#2 pins are routed to the PCH
- BIOS configured to auto-detect if B31 + B81 is low

Pin	Side B Connector		Side A Connector	
#	Name	Description	Name	Description
1	+12v	+12 volt power	PRSNT#1	Hot-plug presence detect
2	+12v	+12 volt power	+12v	+12 volt power
3	+12v	+12 volt power	+12v	+12 volt power
4	GND	Ground	GND	Ground
5	SMCLK	SMBus clock	JTAG2	TCK
6	SMDAT	SMBus data	JTAG3	TDI
7	GND	Ground	JTAG4	TDO
8	+3.3v	+3.3 volt power	JTAG5	TMS
9	JTAG1	+TRST#	+3.3v	+3.3 volt power
10	3.3Vaux	+3.3 volt power	+3.3v	+3.3 volt power
11	WAKE#	Link Reactivation	PWRGD	Power Good
Mechanical Key				
12	SMBALERT#	Power Reduction	GND	Ground
13	GND	Ground	REFCLK+	Reference Clock
14	PETP(0)	Transmitter Lane 0,	REFCLK-	Differential pair
15	PETN(0)	Differential pair	GND	Ground
16	GND	Ground	PERP(0)	Receiver Lane 0,
17	PRSNT#2	Presence detect	PERN(0)	Differential pair
18	GND	Ground	GND	Ground
19	PETP(1)	Transmitter Lane 1,	RSVD	Reserved
20	PETN(1)	Differential pair	GND	Ground
21	GND	Ground	PERP(1)	Receiver Lane 1,
22	GND	Ground	PERN(1)	Differential pair
23	PETP(2)	Transmitter Lane 2,	GND	Ground
24	PETN(2)	Differential pair	GND	Ground
25	GND	Ground	PERP(2)	Receiver Lane 2,
26	GND	Ground	PERN(2)	Differential pair
27	PETP(3)	Transmitter Lane 3,	GND	Ground
28	PETN(3)	Differential pair	GND	Ground
29	GND	Ground	PERP(3)	Receiver Lane 3,
30	PWRBRK#	Power Reduction	PERN(3)	Differential pair
31	BIFURx4	0 = PCIe x4 Bifurcation	GND	Ground

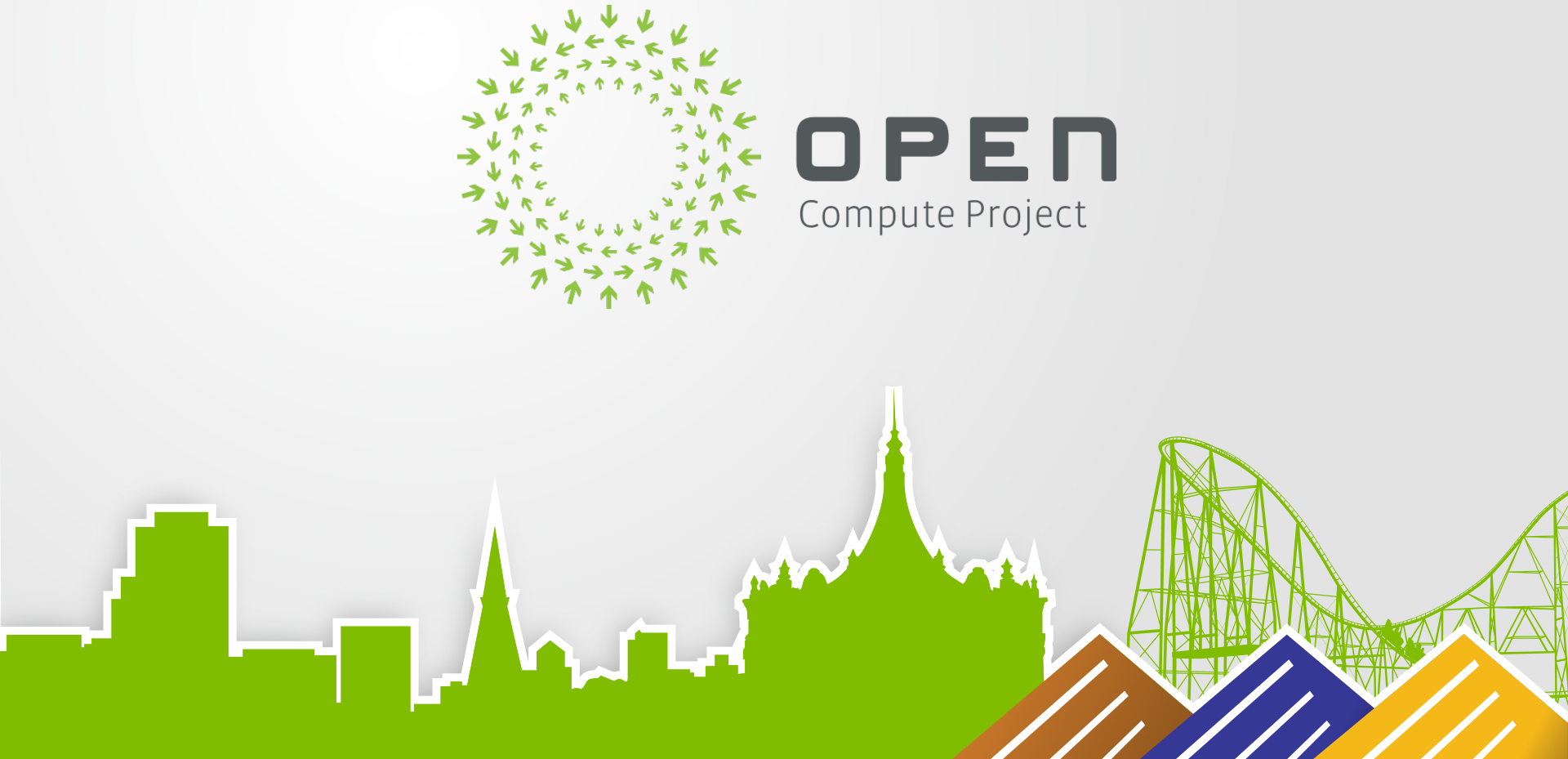
Use cases

- Cache
- Databases
- File system cache



Disaggregate Lab

- AVA and NVME at Disaggregate: Lab
- Partners tested:
  - Excelero
  - Hedvig
  - Weka.IO
  - Spectrum Scale - IBM


OPEN
Compute Project