How to build a GPU cluster

Christopher Milan

AI Safety at UCLA

March, 2025

Who am I

- I am Christopher Milan, the president of AI Safety at UCLA.

Who am I

- I am Christopher Milan, the president of AI Safety at UCLA.
- I spent a significant amount of my time this past year building out our servers.

Who am I

- I am Christopher Milan, the president of AI Safety at UCLA.
- I spent a significant amount of my time this past year building out our servers.
- These slides will be available online at `ais-ucla.org/~chris`.

What is this talk about?

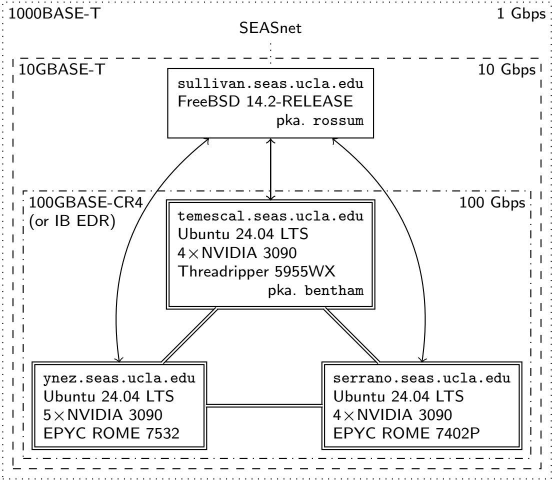
While setting up and administrating our servers, I ran into many interesting challenges.

This talk will focus on how we chose our hardware.

There are many other interesting things to learn from this cluster, check out our blog.[1]

[1] blog.aisafetyatucla.org

Server topology



Why AMD (EPYC ROME)?

AMD's "Zen 2" server CPUs.

Why AMD (EPYC ROME)?

AMD's "Zen 2" server CPUs.

- Memory bandwidth:

Why AMD (EPYC ROME)?

AMD's "Zen 2" server CPUs.

- Memory bandwidth: 8 channels of DDR4-3200, roughly 200GB/s.

Why AMD (EPYC ROME)?

AMD's "Zen 2" server CPUs.

- Memory bandwidth: 8 channels of DDR4-3200, roughly 200GB/s.
- Price:

Why AMD (EPYC ROME)?

AMD's "Zen 2" server CPUs.

- Memory bandwidth: 8 channels of DDR4-3200, roughly 200GB/s.
- Price: A 7532 is about $200 on ebay.

Why AMD (EPYC ROME)?

AMD's "Zen 2" server CPUs.

- Memory bandwidth: 8 channels of DDR4-3200, roughly 200GB/s.
- Price: A 7532 is about $200 on ebay.
- PCIe lanes:

Why AMD (EPYC ROME)?

AMD's "Zen 2" server CPUs.

- Memory bandwidth: 8 channels of DDR4-3200, roughly 200GB/s.
- Price: A 7532 is about $200 on ebay.
- PCIe lanes: EPYC ROME has 128 PCIe 4.0 lanes.

ROMED8-2T

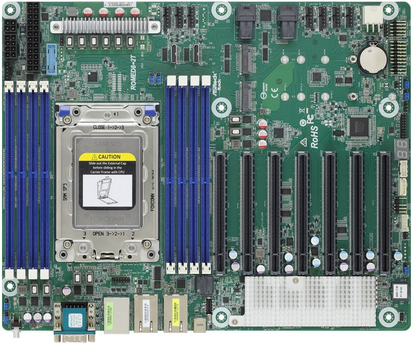


Figure: ASRock Rack ROMED8-2T

- Reasonably priced (at the time...)

ROMED8-2T

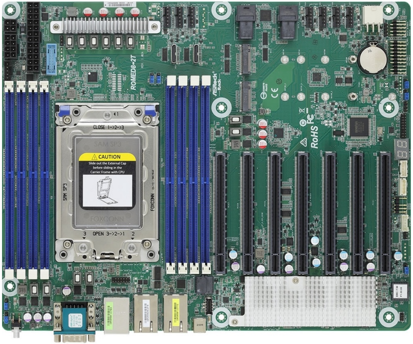


Figure: ASRock Rack ROMED8-2T

- Reasonably priced (at the time...)
- 7×PCIe 4.0x16

ROMED8-2T

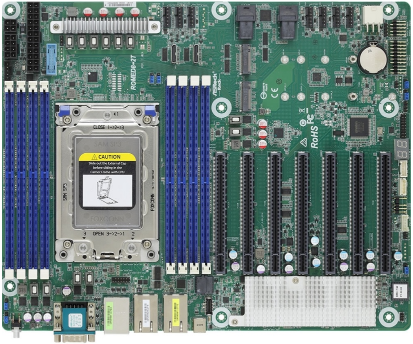


Figure: ASRock Rack ROMED8-2T

- Reasonably priced (at the time...)
- 7×PCIe 4.0x16 (but slot #2 is cursed)

ROMED8-2T

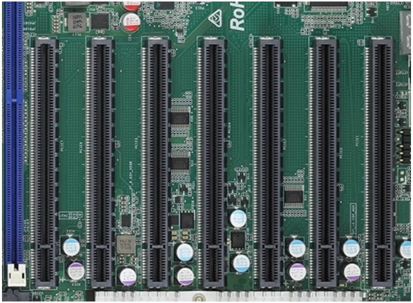


Figure: ROMED8-2T PCIe slots

- Reasonably priced (at the time...)
- 7×PCIe 4.0x16 (but slot #2 is cursed)
- The slots are too close!

PCIe risers to the rescue!

- Crypto miners have this exact same problem

PCIe risers to the rescue!

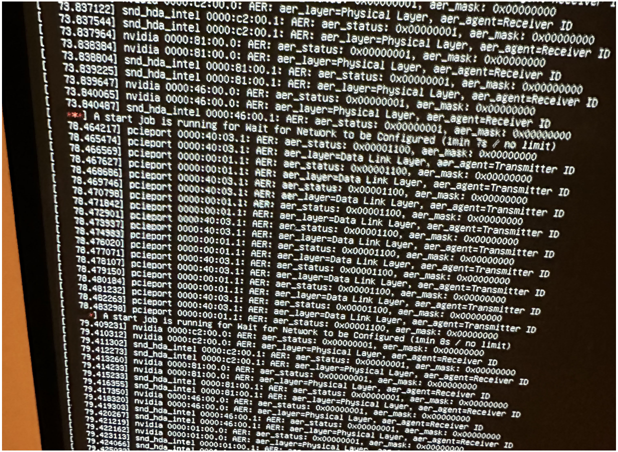
- Crypto miners have this exact same problem
- Solution: PCIe risers

PCIe risers to the rescue!

- Crypto miners have this exact same problem
- Solution: PCIe risers
- Surely nothing could go wrong...

Figure: PCIe risers on `temescal`

Signal integrity errors



```
[   73.837122] snd_hda_intel 0000:c2:00.1: AER: aer_status: 0x00000001, aer_mask: 0x00000000
[   73.837544] snd_hda_intel 0000:c2:00.1: AER: aer_layer=Physical Layer, aer_agent=Receiver ID
[   73.837964] nvidia 0000:81:00.0: AER: aer_status: 0x00000001, aer_mask: 0x00000000
[   73.838384] nvidia 0000:81:00.0: AER: aer_layer=Physical Layer, aer_agent=Receiver ID
[   73.838804] snd_hda_intel 0000:81:00.1: AER: aer_status: 0x00000001, aer_mask: 0x00000000
[   73.839225] snd_hda_intel 0000:81:00.1: AER: aer_layer=Physical Layer, aer_agent=Receiver ID
[   73.839647] nvidia 0000:46:00.0: AER: aer_status: 0x00000001, aer_mask: 0x00000000
[   73.840065] nvidia 0000:46:00.0: AER: aer_layer=Physical Layer, aer_agent=Receiver ID
[   73.840487] snd_hda_intel 0000:46:00.1: AER: aer_status: 0x00000001, aer_mask: 0x00000000
[***] A start job is running for Wait for Network to be Configured (1min 7s / no limit)
[   78.464217] pcieport 0000:40:03.1: AER: aer_status: 0x00001100, aer_mask: 0x00000000
[   78.465474] pcieport 0000:40:03.1: AER: aer_layer=Data Link Layer, aer_agent=Transmitter ID
[   78.466569] pcieport 0000:00:01.1: AER: aer_status: 0x00001100, aer_mask: 0x00000000
[   78.467627] pcieport 0000:00:01.1: AER: aer_layer=Data Link Layer, aer_agent=Transmitter ID
[   78.468686] pcieport 0000:40:03.1: AER: aer_status: 0x00001100, aer_mask: 0x00000000
[   78.469746] pcieport 0000:40:03.1: AER: aer_layer=Data Link Layer, aer_agent=Transmitter ID
[   78.470798] pcieport 0000:00:01.1: AER: aer_status: 0x00001100, aer_mask: 0x00000000
[   78.471842] pcieport 0000:00:01.1: AER: aer_layer=Data Link Layer, aer_agent=Transmitter ID
[   78.472901] pcieport 0000:40:03.1: AER: aer_status: 0x00001100, aer_mask: 0x00000000
[   78.473937] pcieport 0000:40:03.1: AER: aer_layer=Data Link Layer, aer_agent=Transmitter ID
[   78.474983] pcieport 0000:00:01.1: AER: aer_status: 0x00001100, aer_mask: 0x00000000
[   78.476020] pcieport 0000:00:01.1: AER: aer_layer=Data Link Layer, aer_agent=Transmitter ID
[   78.477071] pcieport 0000:40:03.1: AER: aer_status: 0x00001100, aer_mask: 0x00000000
[   78.478107] pcieport 0000:40:03.1: AER: aer_layer=Data Link Layer, aer_agent=Transmitter ID
[   78.479150] pcieport 0000:00:01.1: AER: aer_status: 0x00001100, aer_mask: 0x00000000
[   78.480184] pcieport 0000:00:01.1: AER: aer_layer=Data Link Layer, aer_agent=Transmitter ID
[   78.481232] pcieport 0000:40:03.1: AER: aer_status: 0x00001100, aer_mask: 0x00000000
[   78.482263] pcieport 0000:40:03.1: AER: aer_layer=Data Link Layer, aer_agent=Transmitter ID
[   78.483298] pcieport 0000:00:01.1: AER: aer_status: 0x00001100, aer_mask: 0x00000000
[ *] A start job is running for Wait for Network to be Configured (1min 8s / no limit)
[   79.409231] nvidia 0000:c2:00.0: AER: aer_status: 0x00000001, aer_mask: 0x00000000
[   79.410312] nvidia 0000:c2:00.0: AER: aer_layer=Physical Layer, aer_agent=Receiver ID
[   79.411302] snd_hda_intel 0000:c2:00.1: AER: aer_status: 0x00000001, aer_mask: 0x00000000
[   79.412273] snd_hda_intel 0000:c2:00.1: AER: aer_layer=Physical Layer, aer_agent=Receiver ID
[   79.413260] nvidia 0000:81:00.0: AER: aer_status: 0x00000001, aer_mask: 0x00000000
[   79.414233] nvidia 0000:81:00.0: AER: aer_layer=Physical Layer, aer_agent=Receiver ID
[   79.415233] snd_hda_intel 0000:81:00.1: AER: aer_status: 0x00000001, aer_mask: 0x00000000
[   79.416355] snd_hda_intel 0000:81:00.1: AER: aer_layer=Physical Layer, aer_agent=Receiver ID
[   79.417350] nvidia 0000:46:00.0: AER: aer_status: 0x00000001, aer_mask: 0x00000000
[   79.418320] nvidia 0000:46:00.0: AER: aer_layer=Physical Layer, aer_agent=Receiver ID
[   79.419303] snd_hda_intel 0000:46:00.1: AER: aer_status: 0x00000001, aer_mask: 0x00000000
[   79.420267] snd_hda_intel 0000:46:00.1: AER: aer_layer=Physical Layer, aer_agent=Receiver ID
[   79.421219] nvidia 0000:01:00.0: AER: aer_status: 0x00000001, aer_mask: 0x00000000
[   79.422162] nvidia 0000:01:00.0: AER: aer_layer=Physical Layer, aer_agent=Receiver ID
[   79.423113] snd_hda_intel 0000:01:00.1: AER: aer_status: 0x00000001, aer_mask: 0x00000000
[   79.424066] snd_hda_intel 0000:01:00.1: AER: aer_layer=Physical Layer,
```

Figure: AER errors, image from Nathan Odle

Signal integrity errors

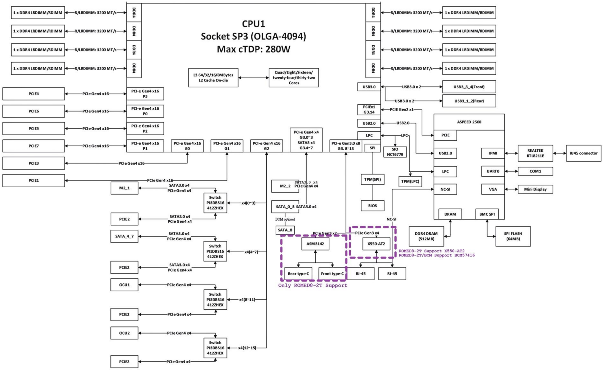


Figure: ROMED8-2T block diagram

Signal integrity errors

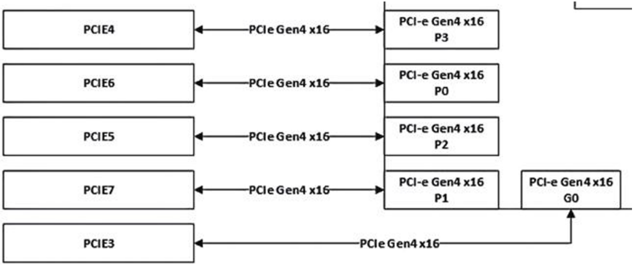


Figure: PCIe slots on ROMED8-2T

Signal integrity errors

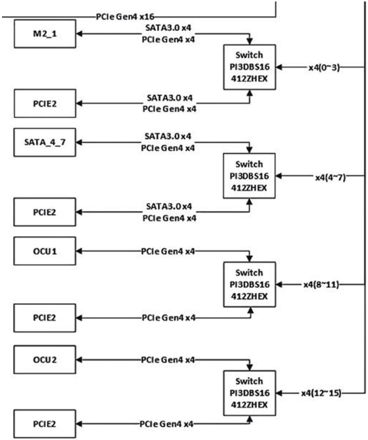


Figure: Remember slot 2? This is him now

Signal integrity errors

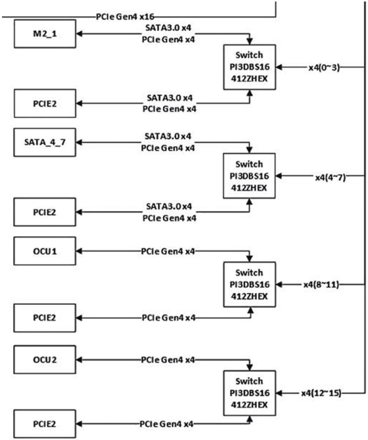


Figure: Remember slot 2? This is him now

All this switching causes signal loss!

Similarly, adding risers does too.

Another issue with PCIe risers

PCIe devices draw power directly from the PSU AND from the PCIe bus itself.

Another issue with PCIe risers

PCIe devices draw power directly from the PSU AND from the PCIe bus itself.

Why is this a problem?

Another issue with PCIe risers

PCIe devices draw power directly from the PSU AND from the PCIe bus itself.

Why is this a problem? Because at 350W/ea. one power supply is not enough for 6 GPUs.

Another issue with PCIe risers

PCIe devices draw power directly from the PSU AND from the PCIe bus itself.

Why is this a problem? Because at 350W/ea. one power supply is not enough for 6 GPUs.

Taping off the PCIe slot power just causes the cards to fail to be recognized.

How to fix this?

Some options:

1. Just ignore it and have fewer GPUs.

How to fix this?

Some options:

1. Just ignore it and have fewer GPUs.
2. Just get better risers, and still have fewer GPUs.

How to fix this?

Some options:

1. Just ignore it and have fewer GPUs.
2. Just get better risers, and still have fewer GPUs.
3. How do the pros do it?

How to fix this?

Some options:

1. Just ignore it and have fewer GPUs.
2. Just get better risers, and still have fewer GPUs.
3. How do the pros do it?

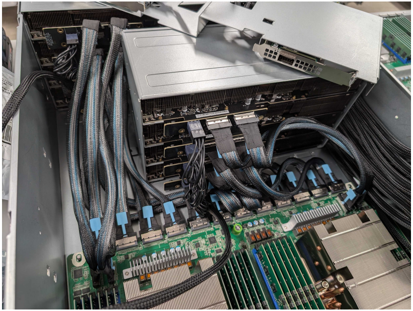


Figure: SlimSAS/MCIO connectors

Dartmouth Summer Research Project on AI

Dartmouth Summer Research Project on AI

We propose [a study] to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it.

Dartmouth Summer Research Project on AI

We propose [a study] to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it.

An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves.

Dartmouth Summer Research Project on AI

> We propose [a study] to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it.
>
> An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves.
>
> We think that a significant advance can be made [. . .] if a carefully selected group of scientists work together for a summer.

Dartmouth Summer Research Project on AI

> We propose [a study] to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it.
>
> An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves.
>
> We think that a significant advance can be made [. . .] if a carefully selected group of scientists work together for a summer.
>
> — A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955

Dartmouth Summer Research Project on AI

> We propose [a study] to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it.
>
> An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves.
>
> We think that a significant advance can be made [. . .] if a carefully selected group of scientists work together for a summer.
>
> — A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955

Solving problems requires tons of repeated trial and error.

The alignment problem

For a sufficiently advanced system, ensuring its goals are the same as ours is a non-trivial task (see universal paperclips).

The alignment problem

For a sufficiently advanced system, ensuring its goals are the same as ours is a non-trivial task (see universal paperclips).

It stands to reason that alignment will be similarly difficult.

The alignment problem

For a sufficiently advanced system, ensuring its goals are the same as ours is a non-trivial task (see universal paperclips).

It stands to reason that alignment will be similarly difficult.

However, we only get one try.

Join us

Visit our website at ais-ucla.org.

Some programs we run:

- Intro to AI Safety Fellowship (ais-ucla.org/fellowships).
- Upskilling Tracks (ais-ucla.org/upskilling-tracks).
- Reading Group (ask Will).
- Server access (ask me).