Cloud-Native Geospatial Data Formats

Christine Smit

NASA / Telophase

Who am I?

B.S. Computer Engineering / Ph.D. Electrical Engineering
focus on audio in research

↓

Started working in Earth science 13 years ago as a software engineer at a NASA archive:
focus on data processing and visualization

↓

Started working in cloud computing 6 years ago (still at NASA):
focus on cloud-based serverless architecture and data processing

Outline

- What terminology is used for data formats in the cloud?
- Where can I get great information?
- How is the cloud different?
- How do you design a data format for the cloud?
- What data formats are out there right now?

What terminology is used for data formats in the cloud?

cloud native ↔ cloud optimized ↔ cloud friendly

What terminology is used for data formats in the cloud?

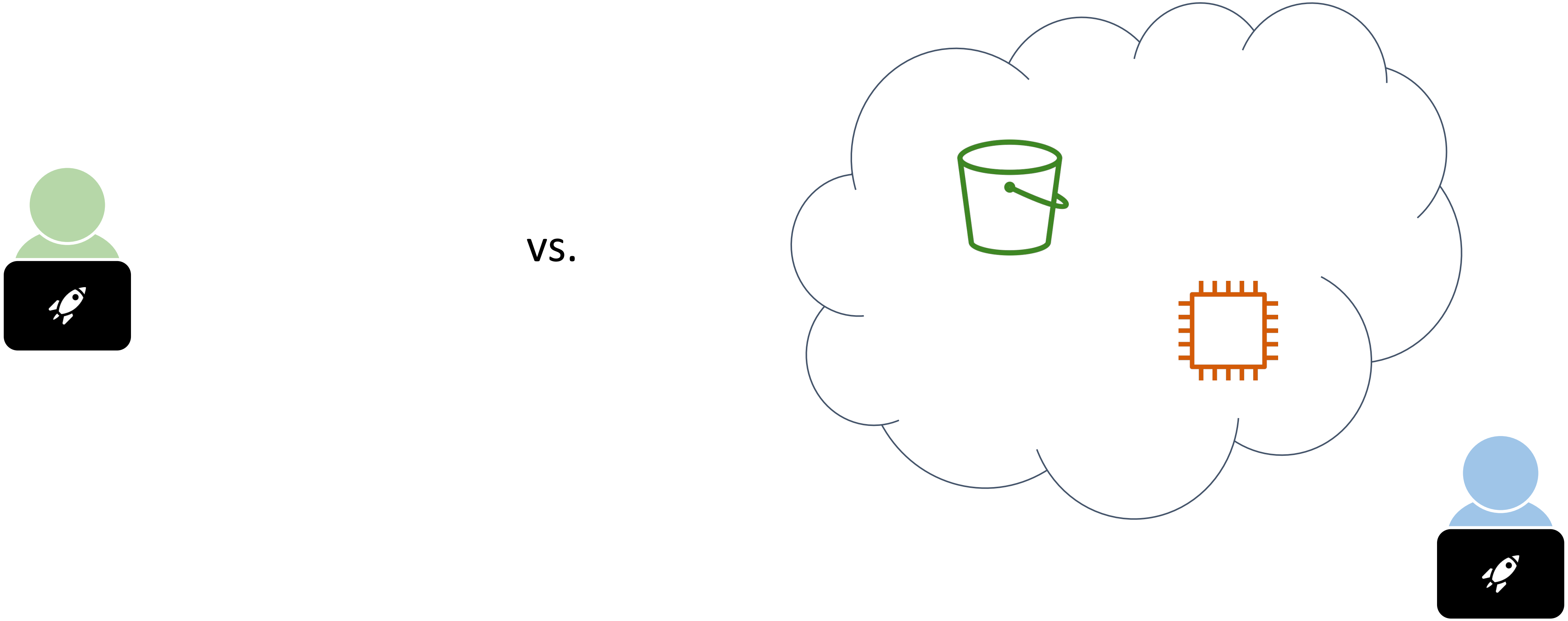
cloud native ↔ cloud optimized ↔ cloud friendly

"ARCO" : analysis ready, cloud optimized

Where can I get great information?

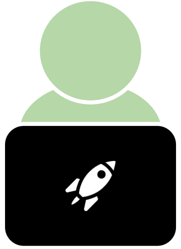
https://guide.cloudnativegeo.org/

How is the cloud different?

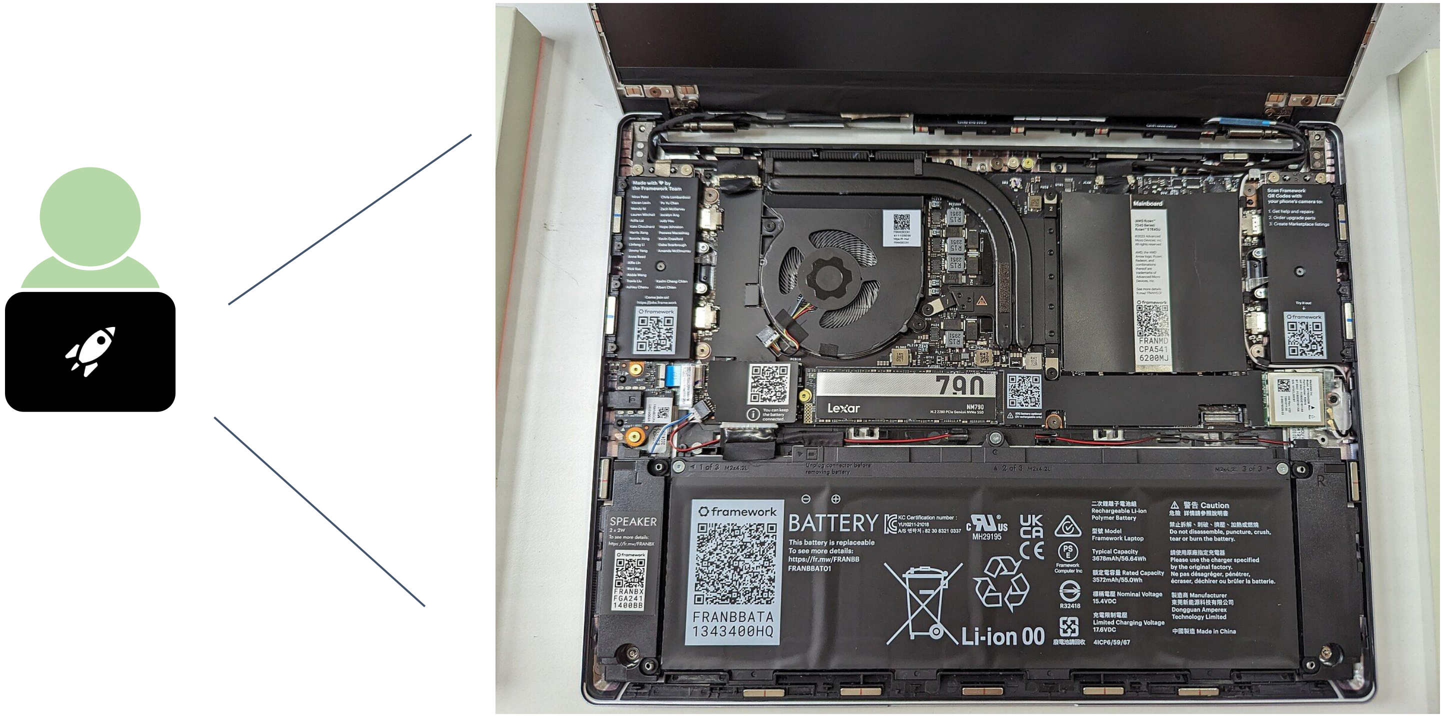


vs.

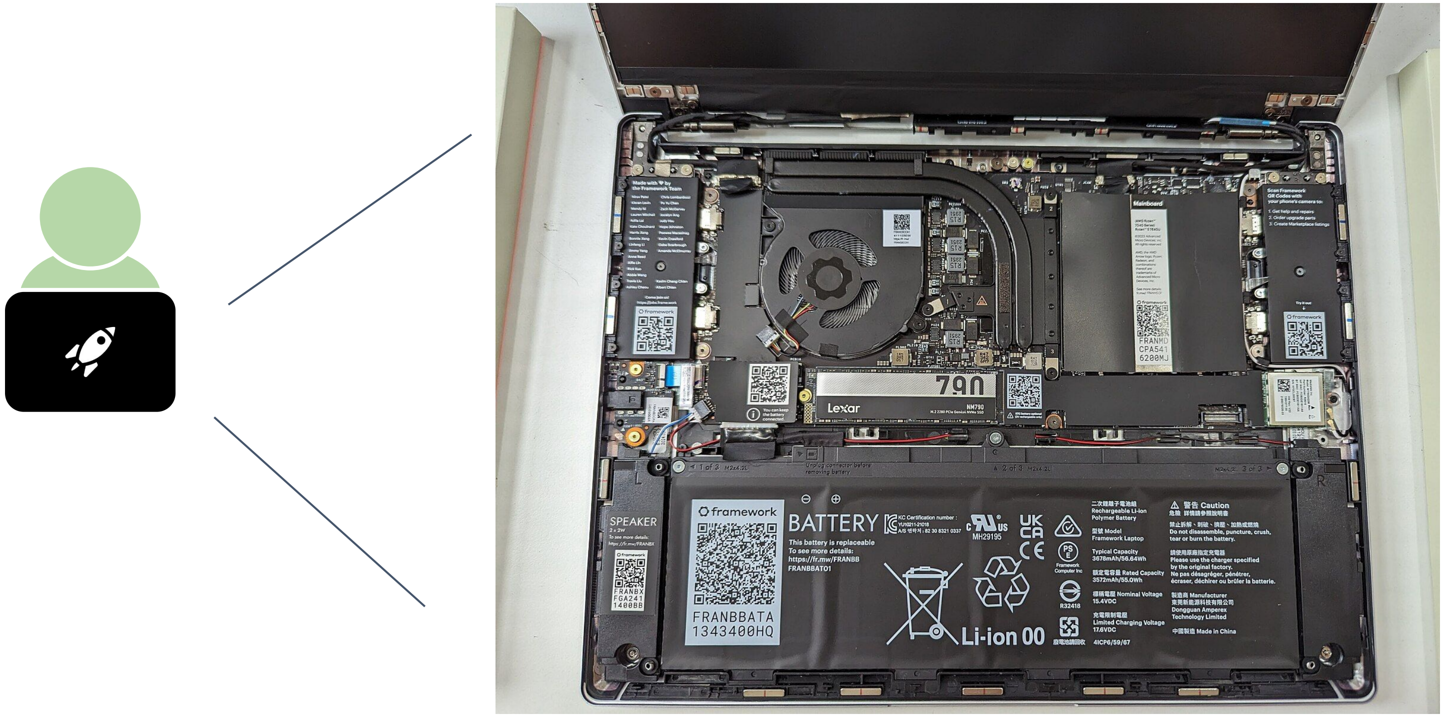
How is the cloud different?



How is the cloud different?

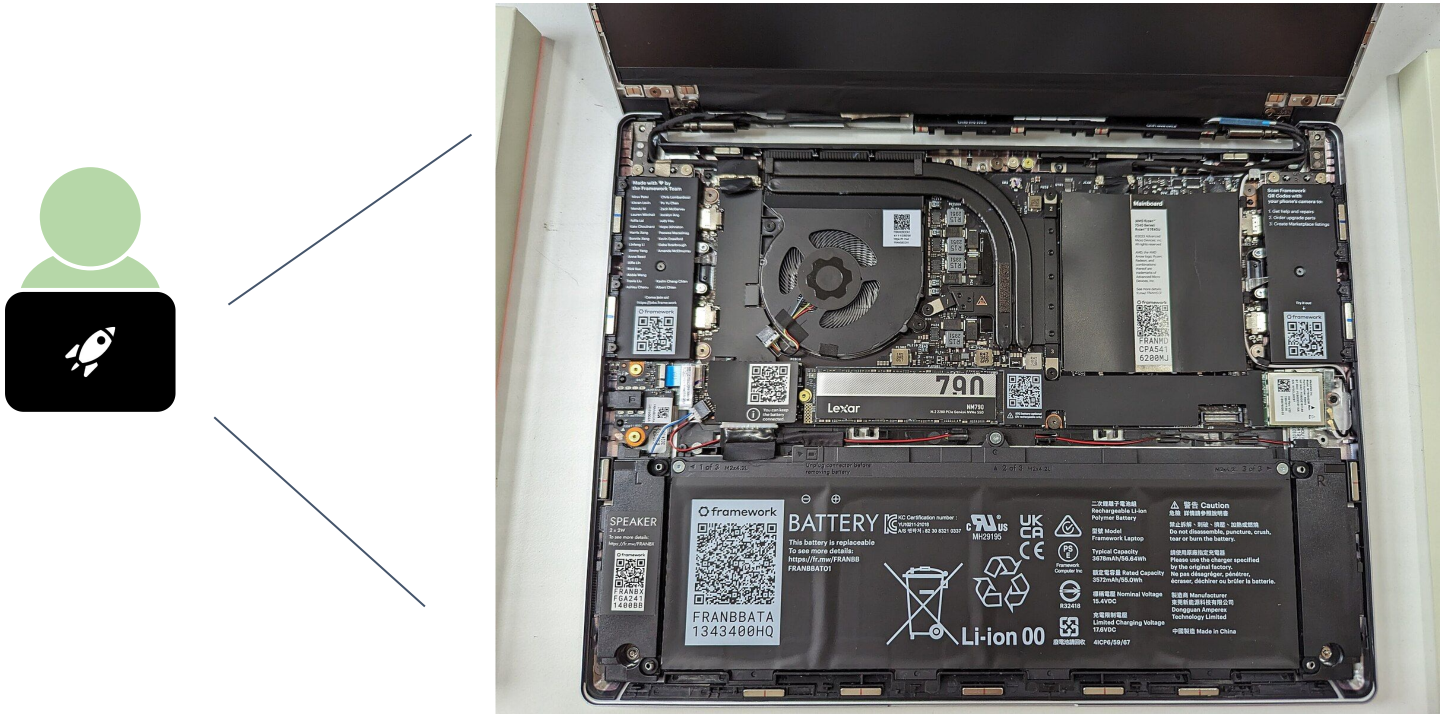


How is the cloud different?



- + Compute and data physically close together (= FAST Communication!!!)

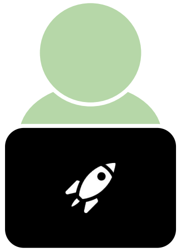
How is the cloud different?



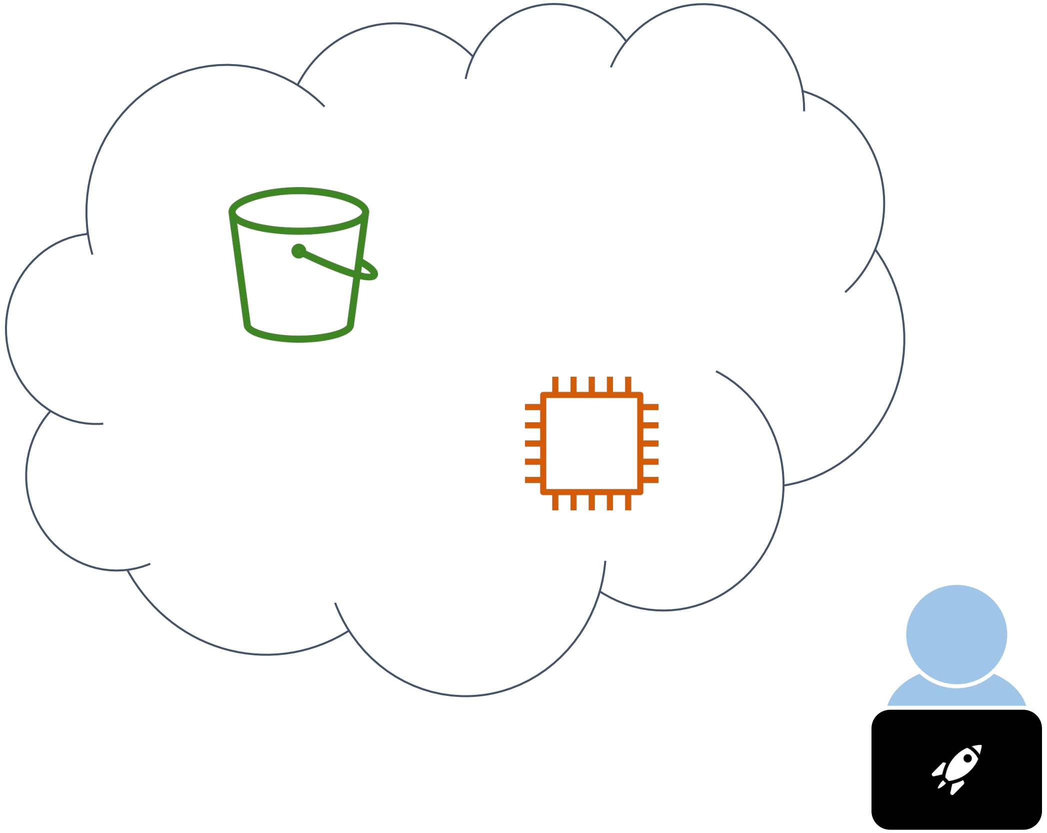
+ Compute and data physically close together (= FAST Communication!!!)

- limited storage
- limited compute

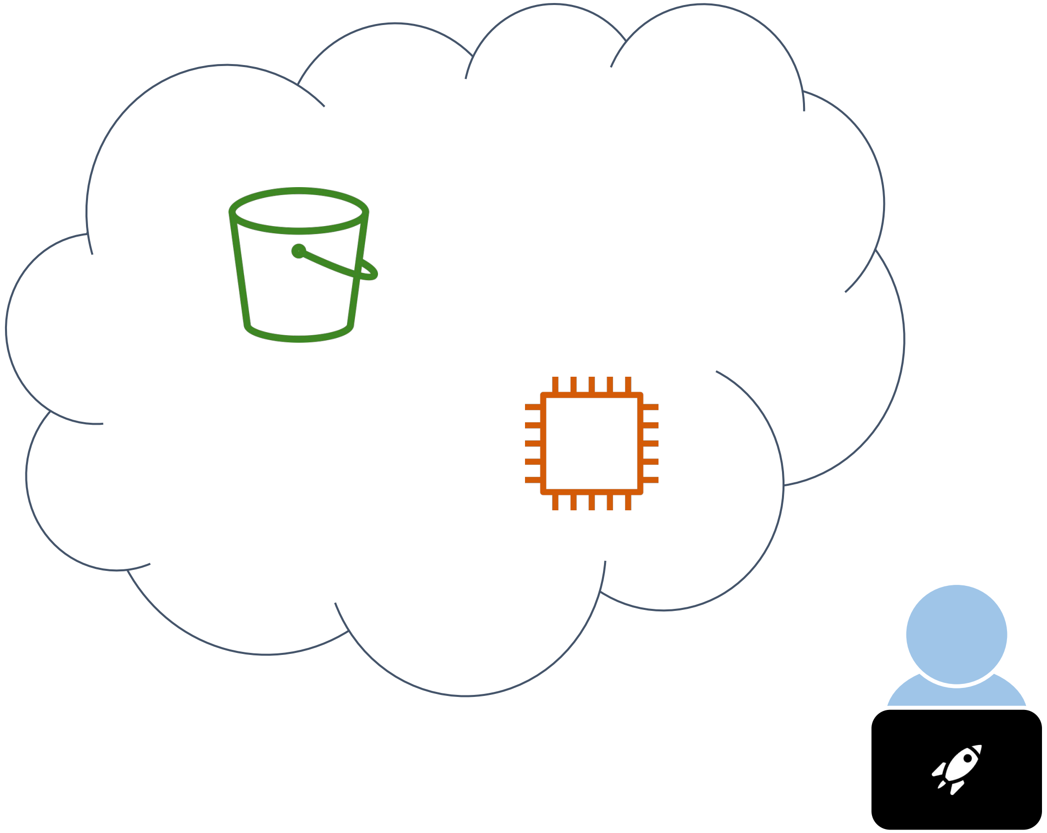
How is the cloud different?



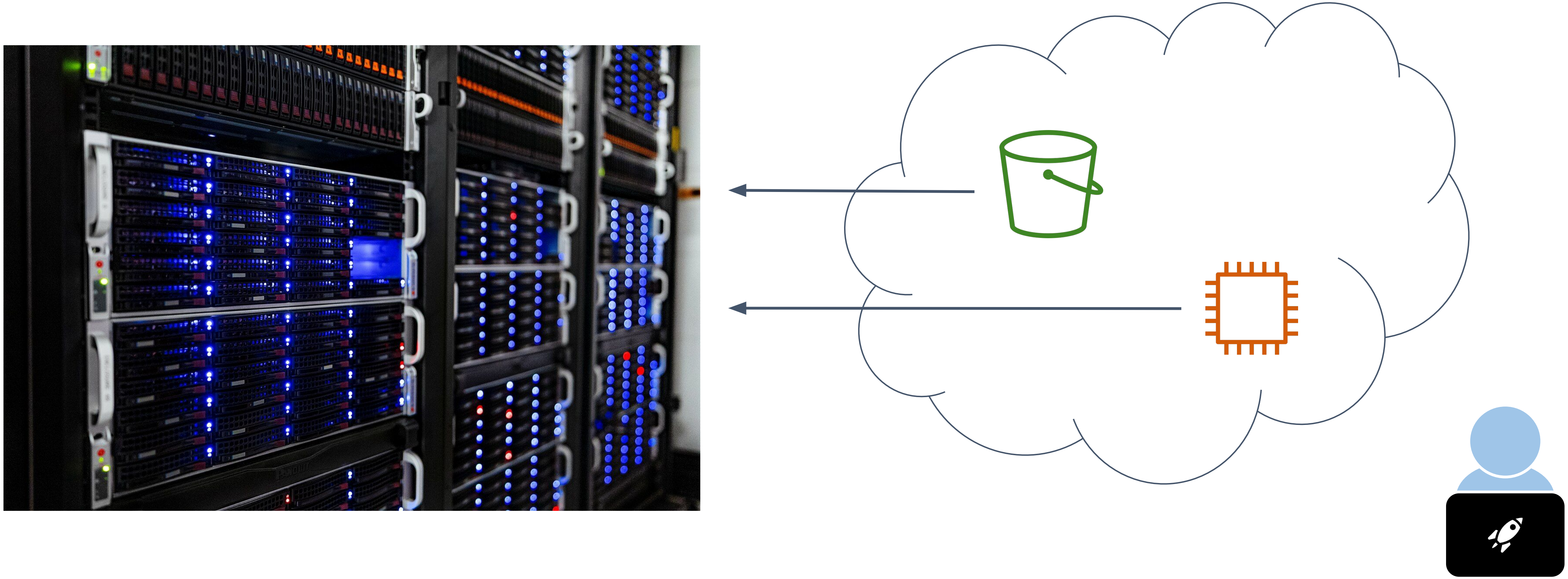
vs.



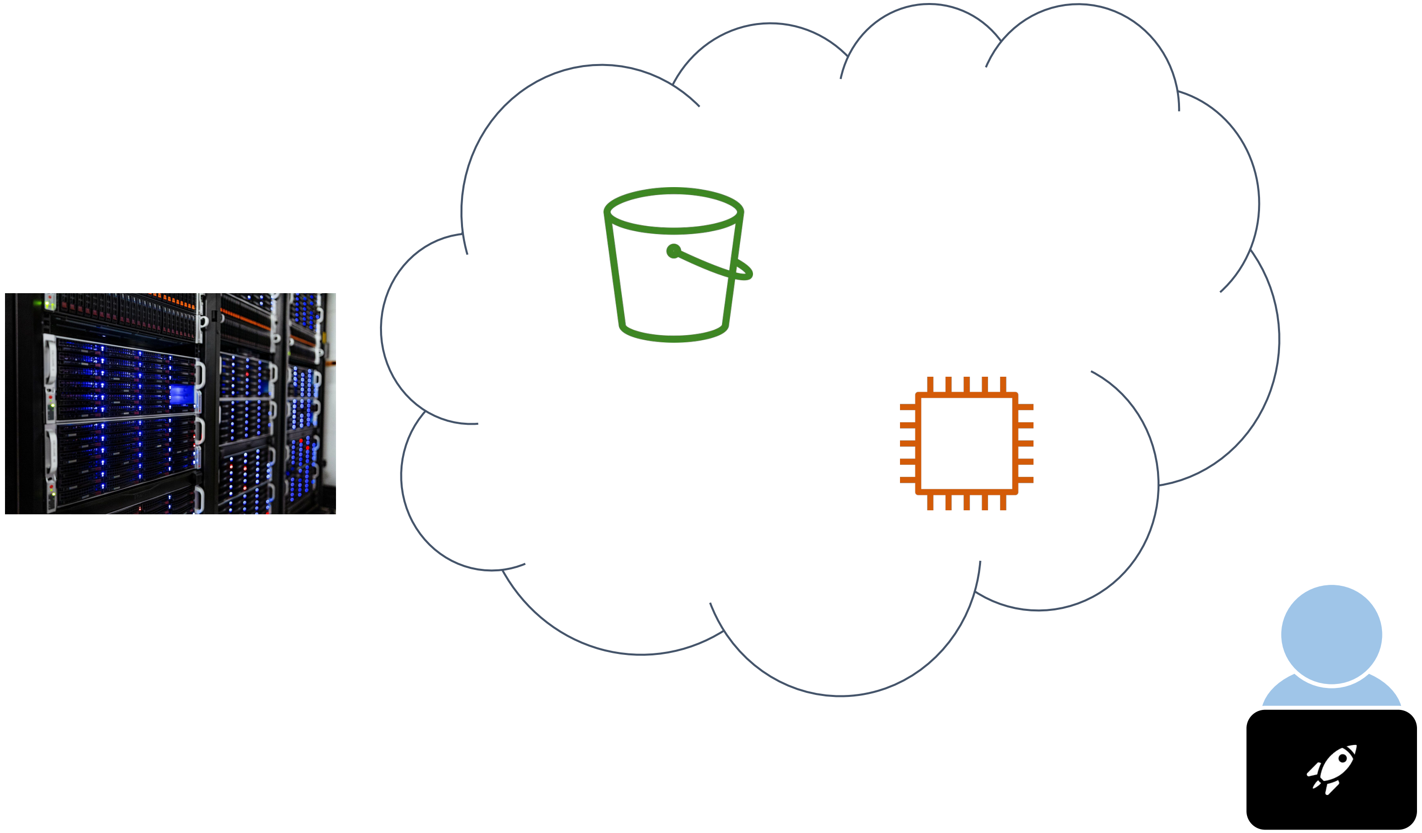
How is the cloud different?



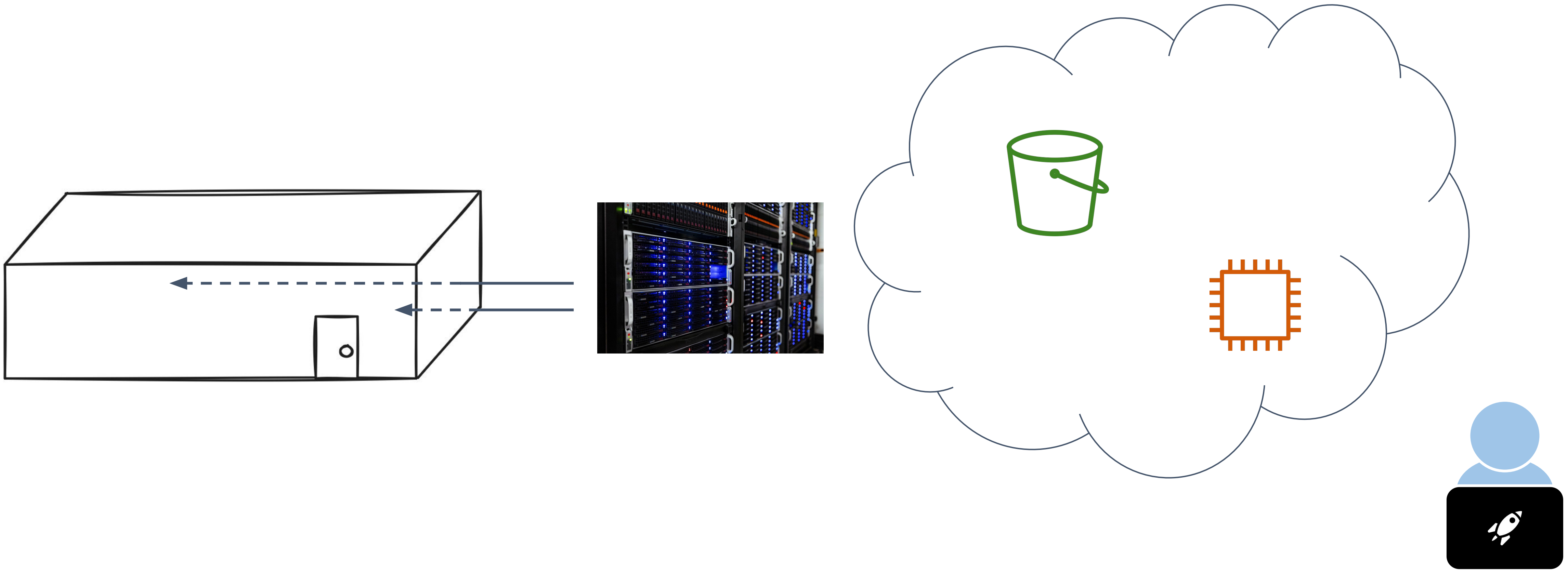
How is the cloud different?



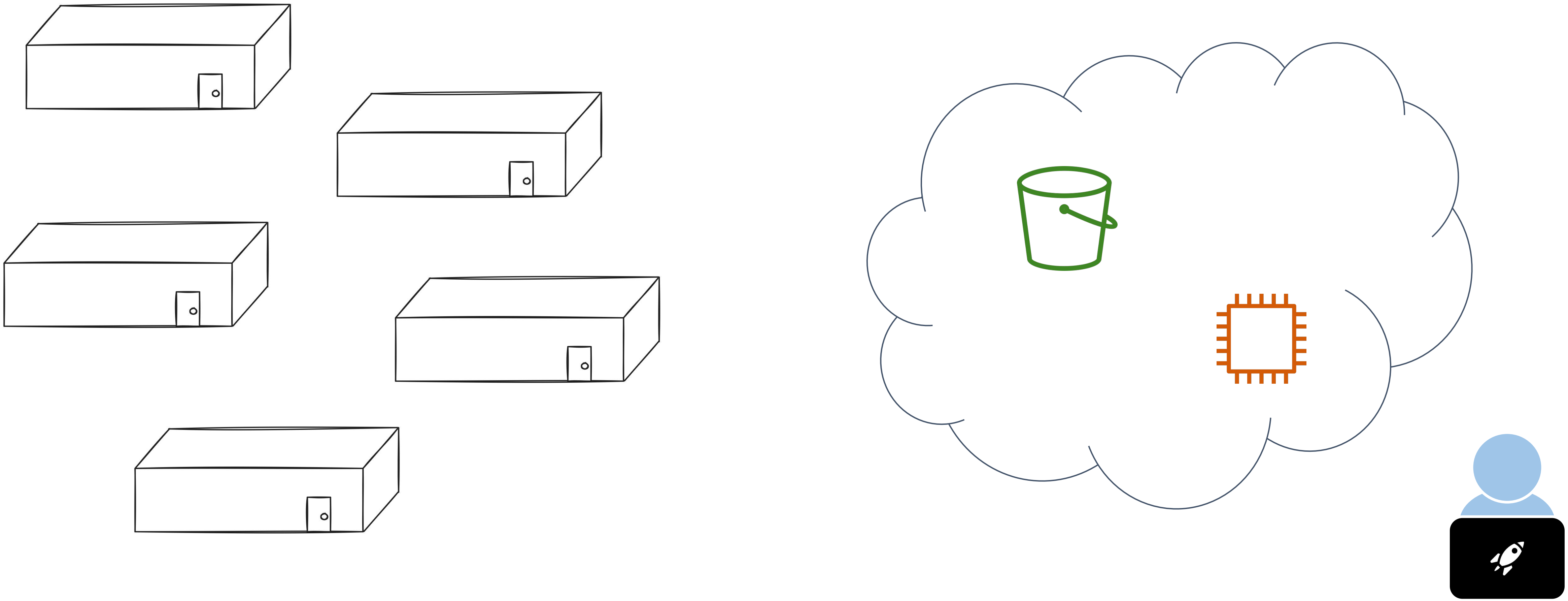
How is the cloud different?



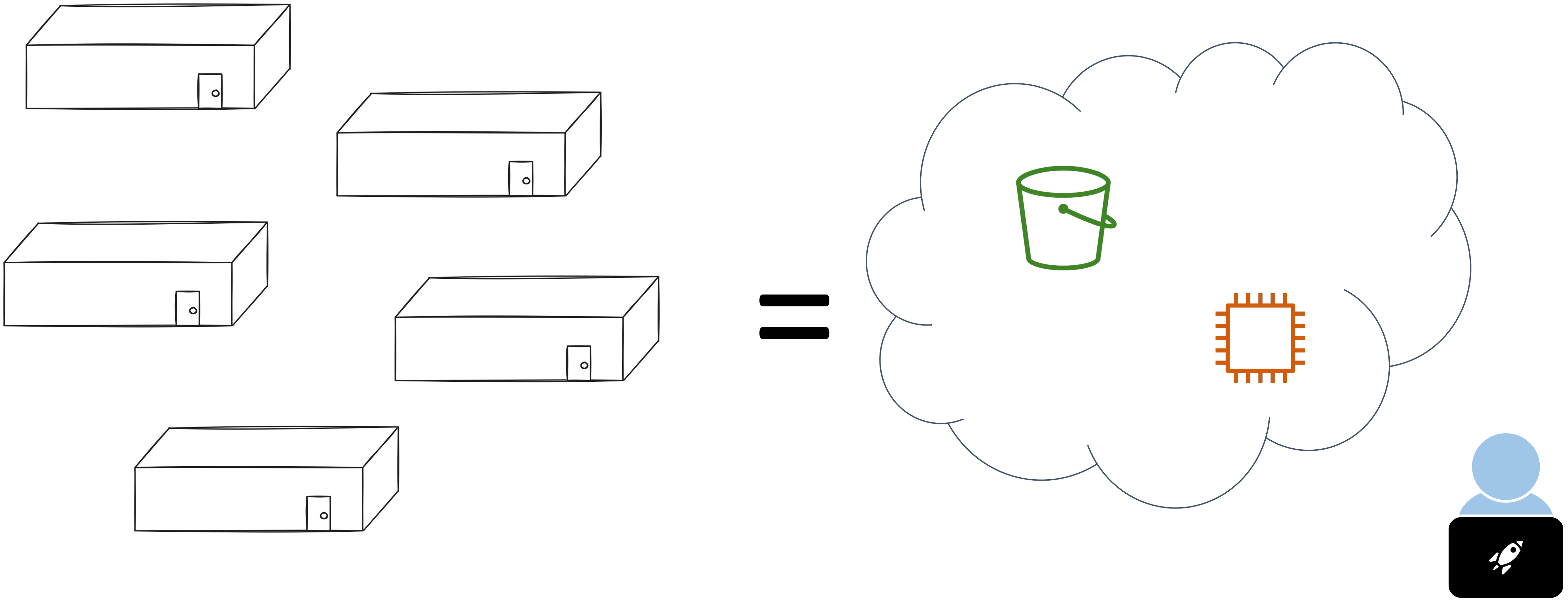
How is the cloud different?



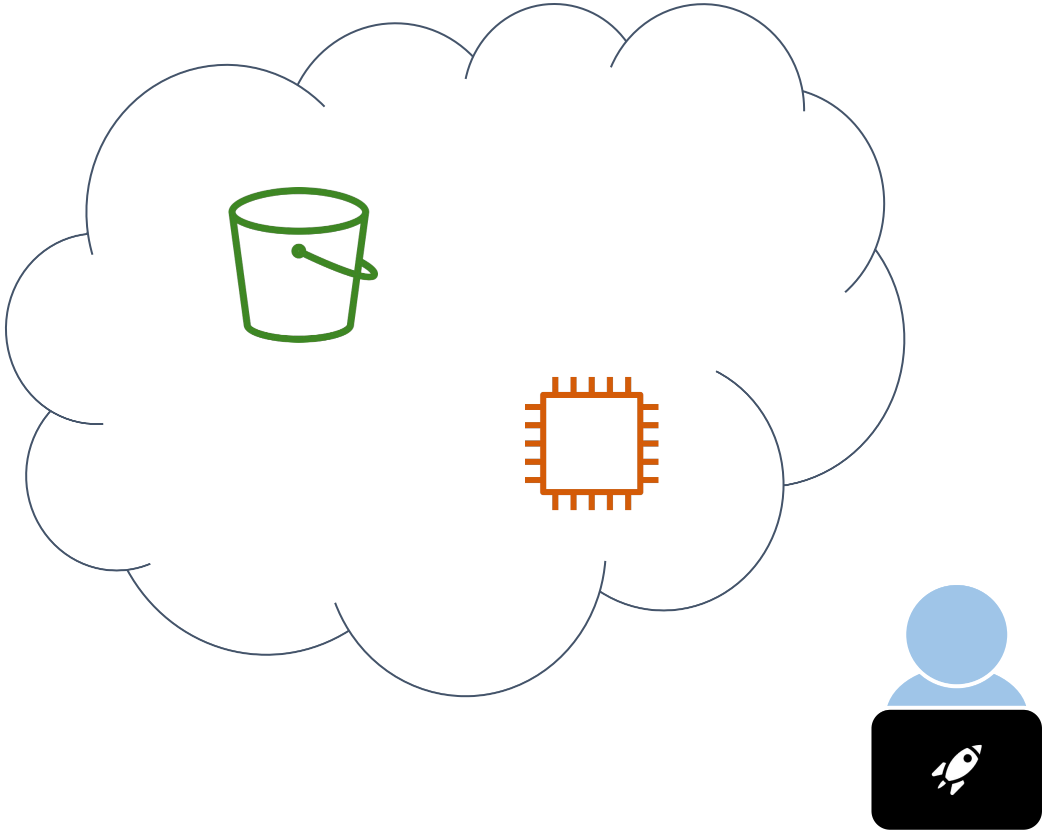
How is the cloud different?



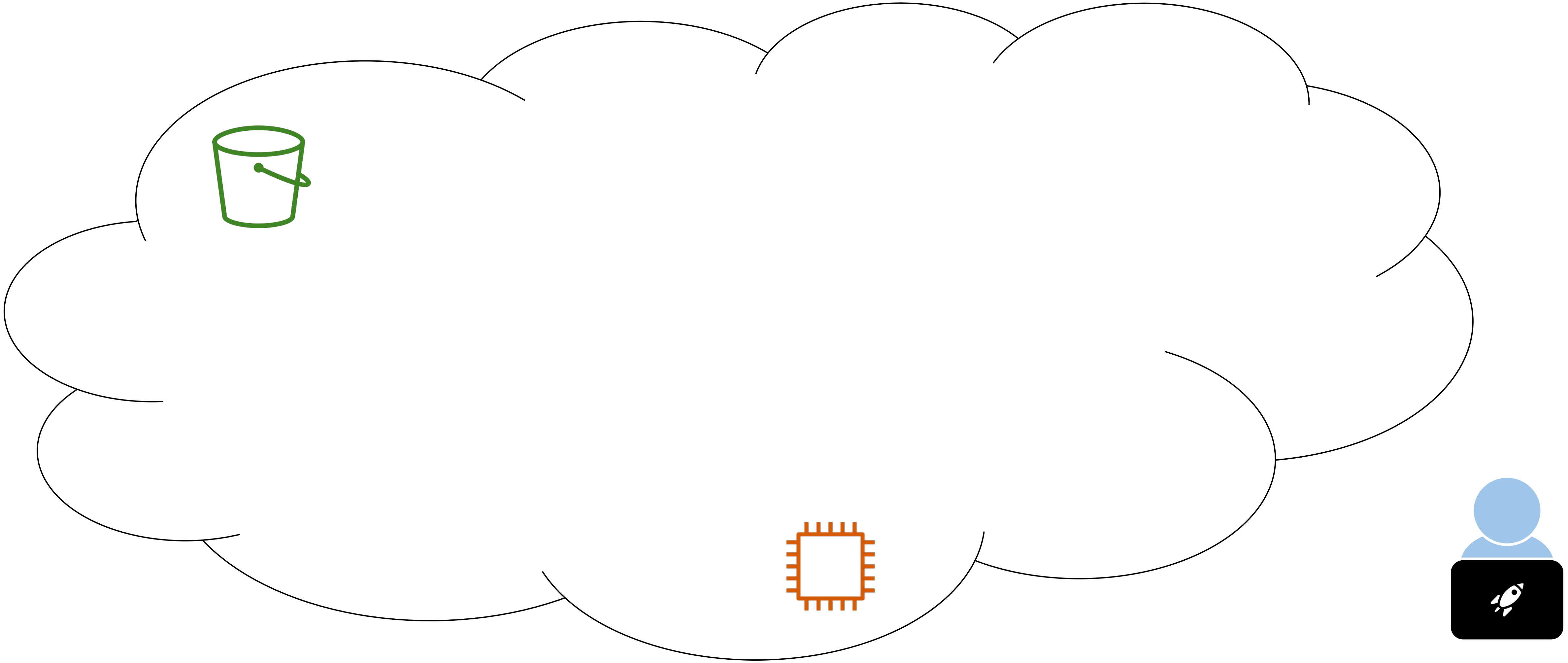
How is the cloud different?



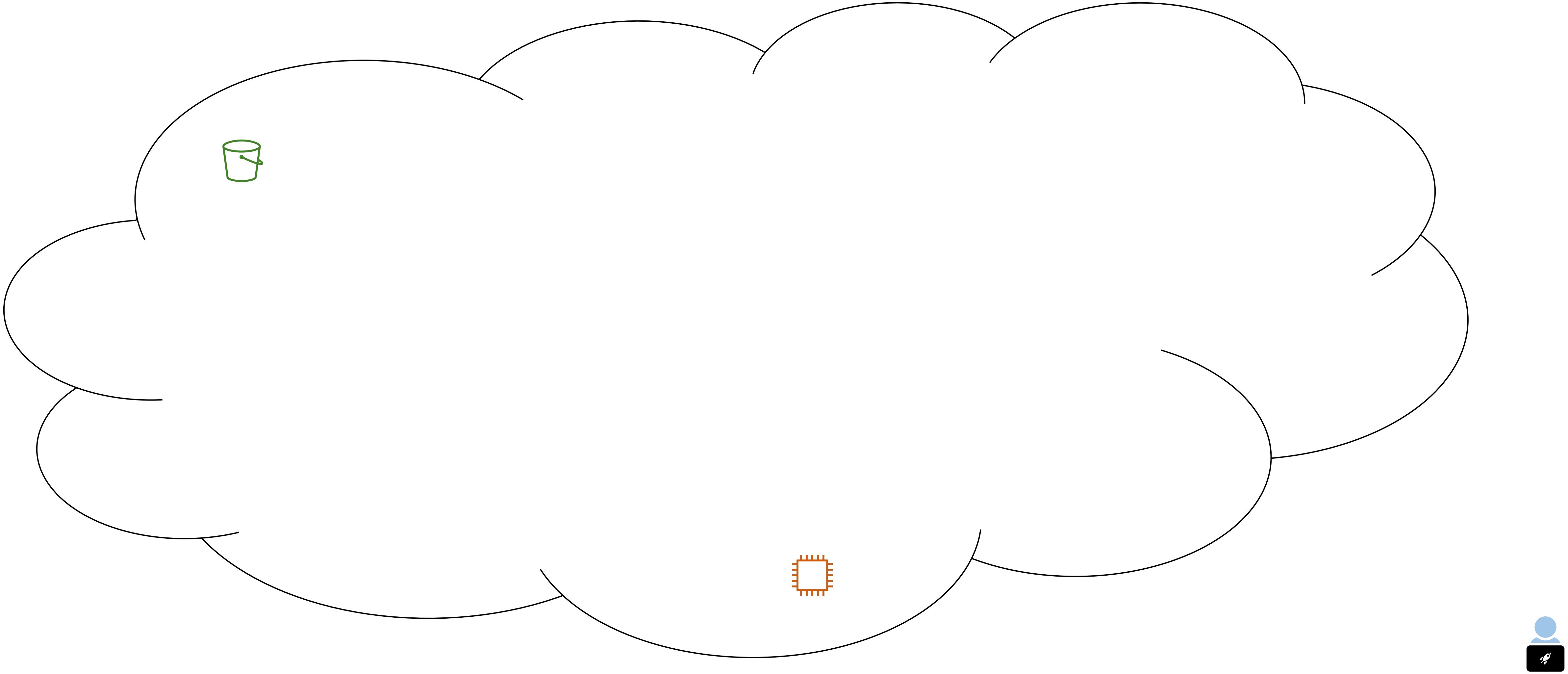
How is the cloud different?



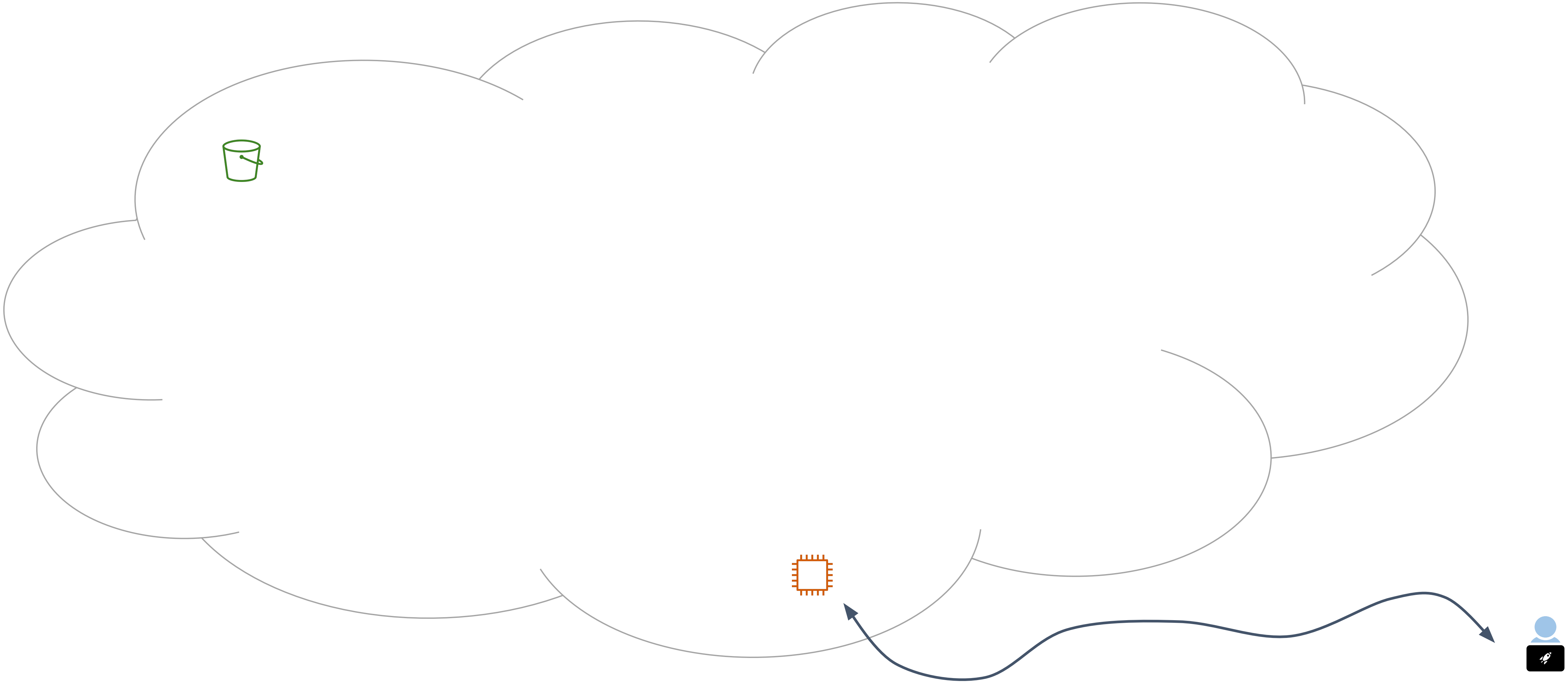
How is the cloud different?



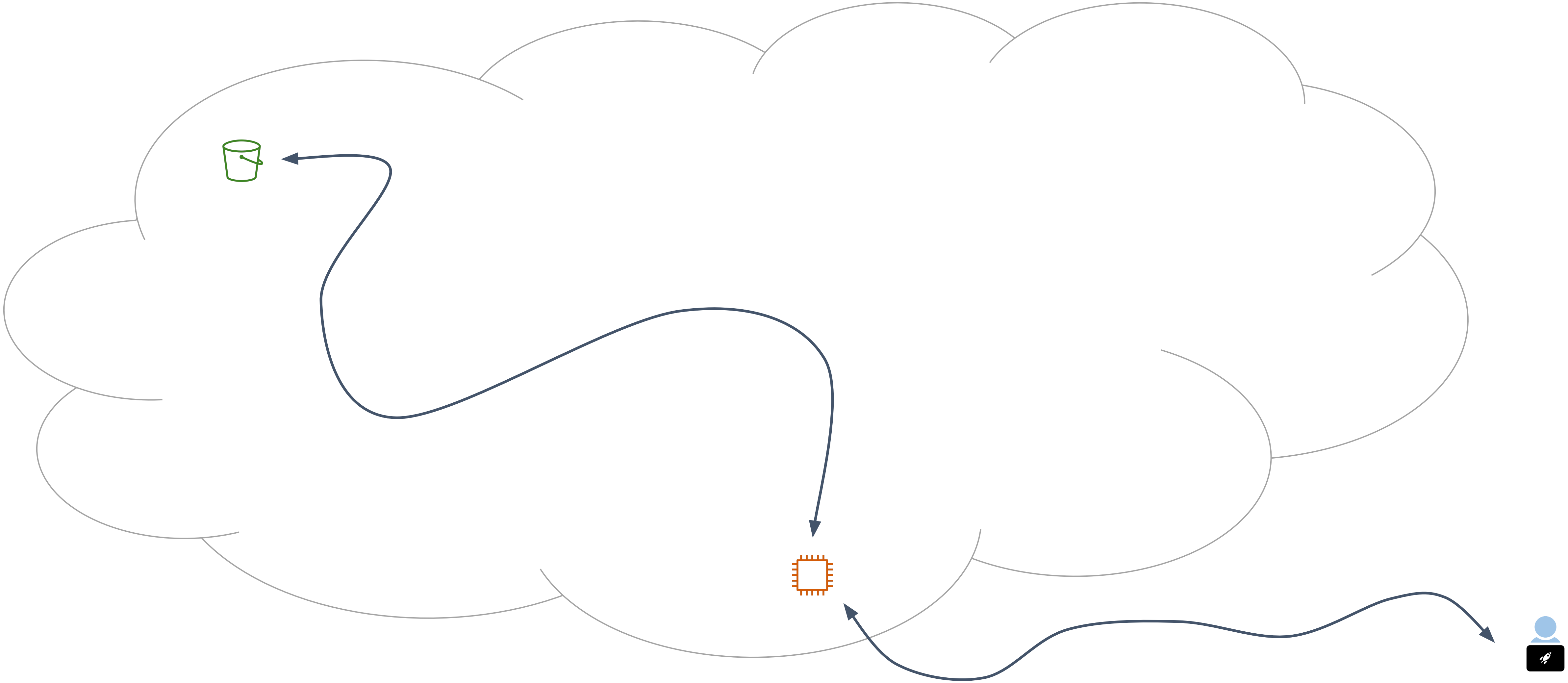
How is the cloud different?



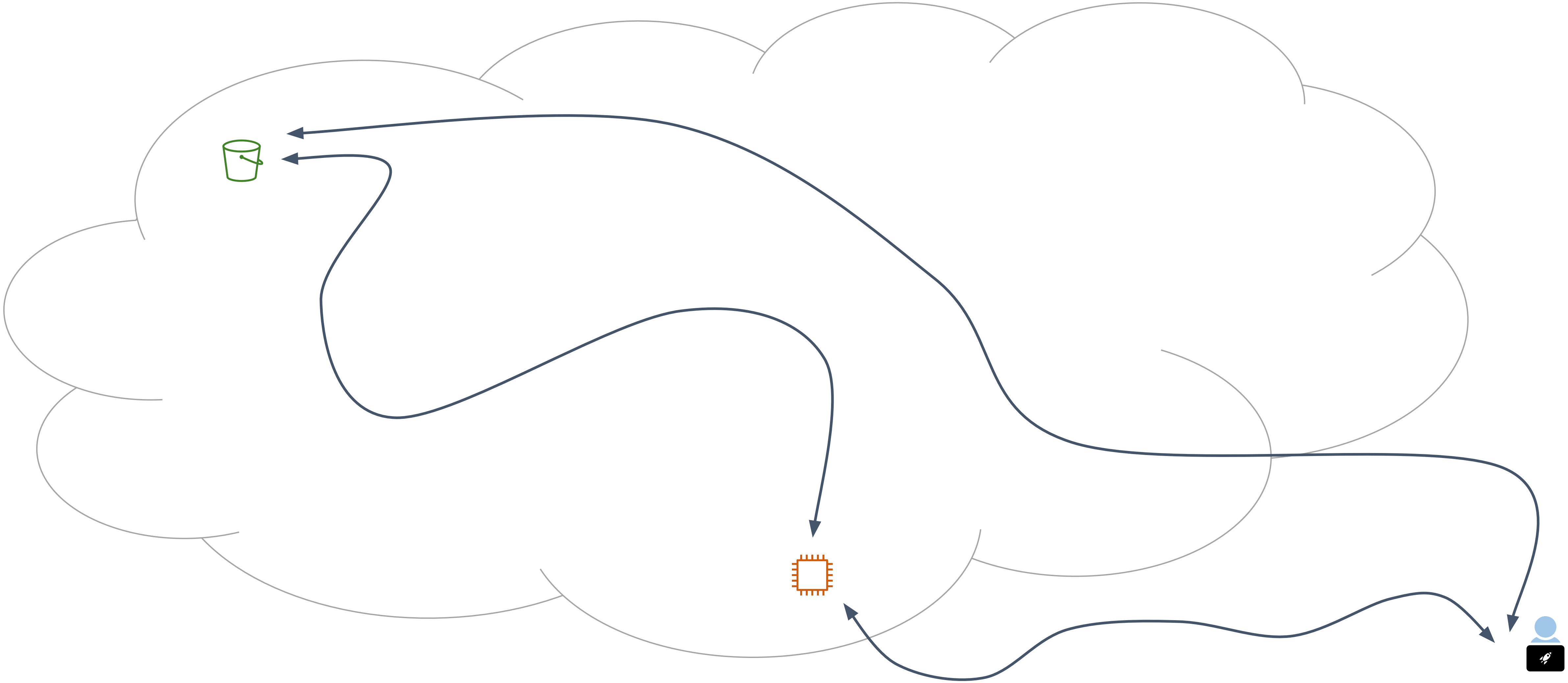
How is the cloud different?



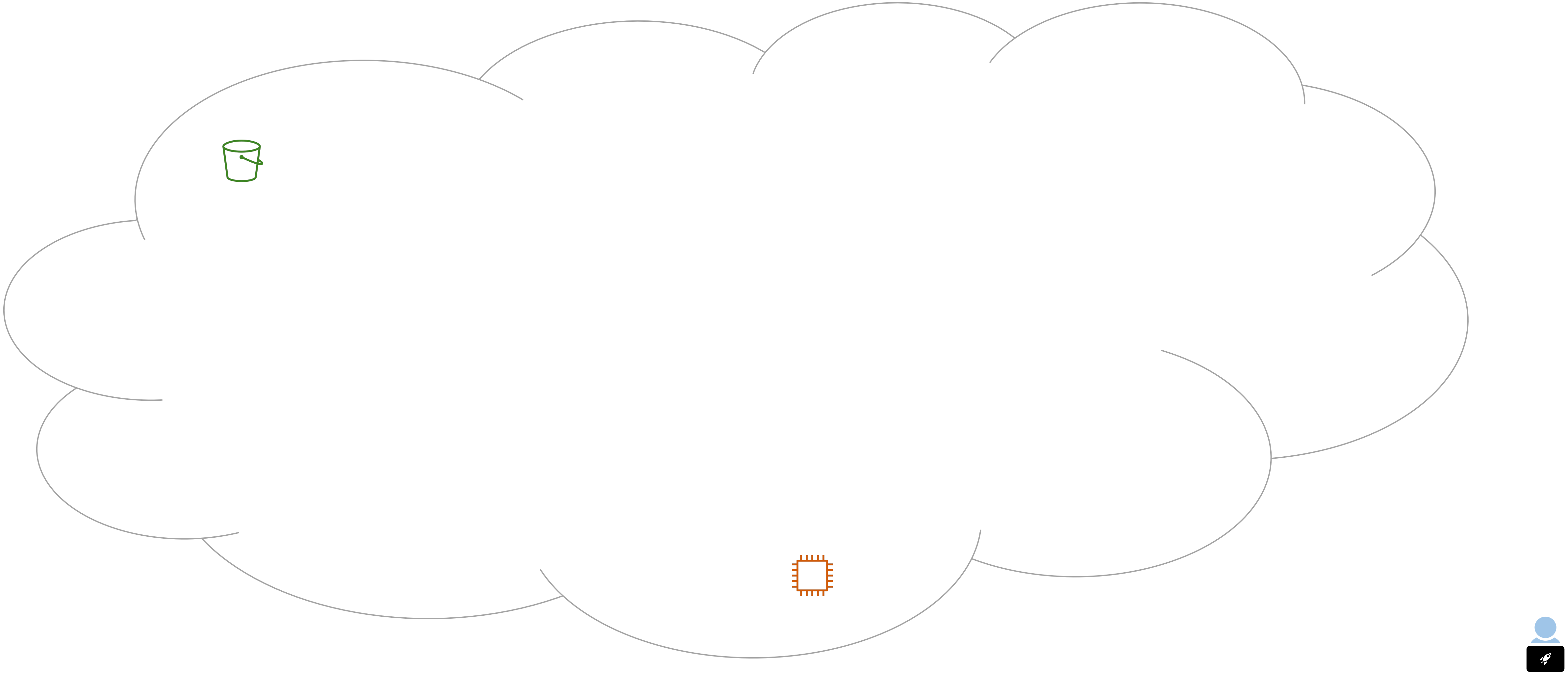
How is the cloud different?



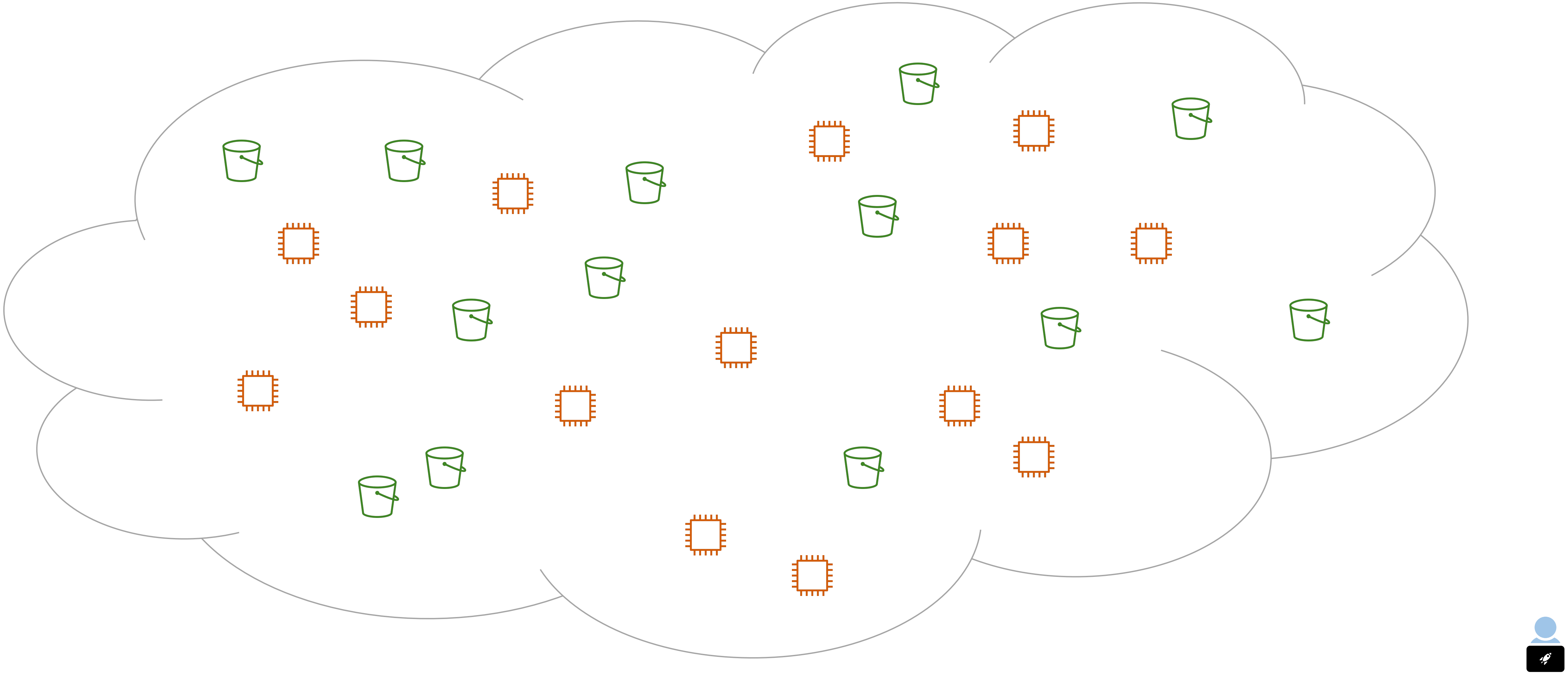
How is the cloud different?



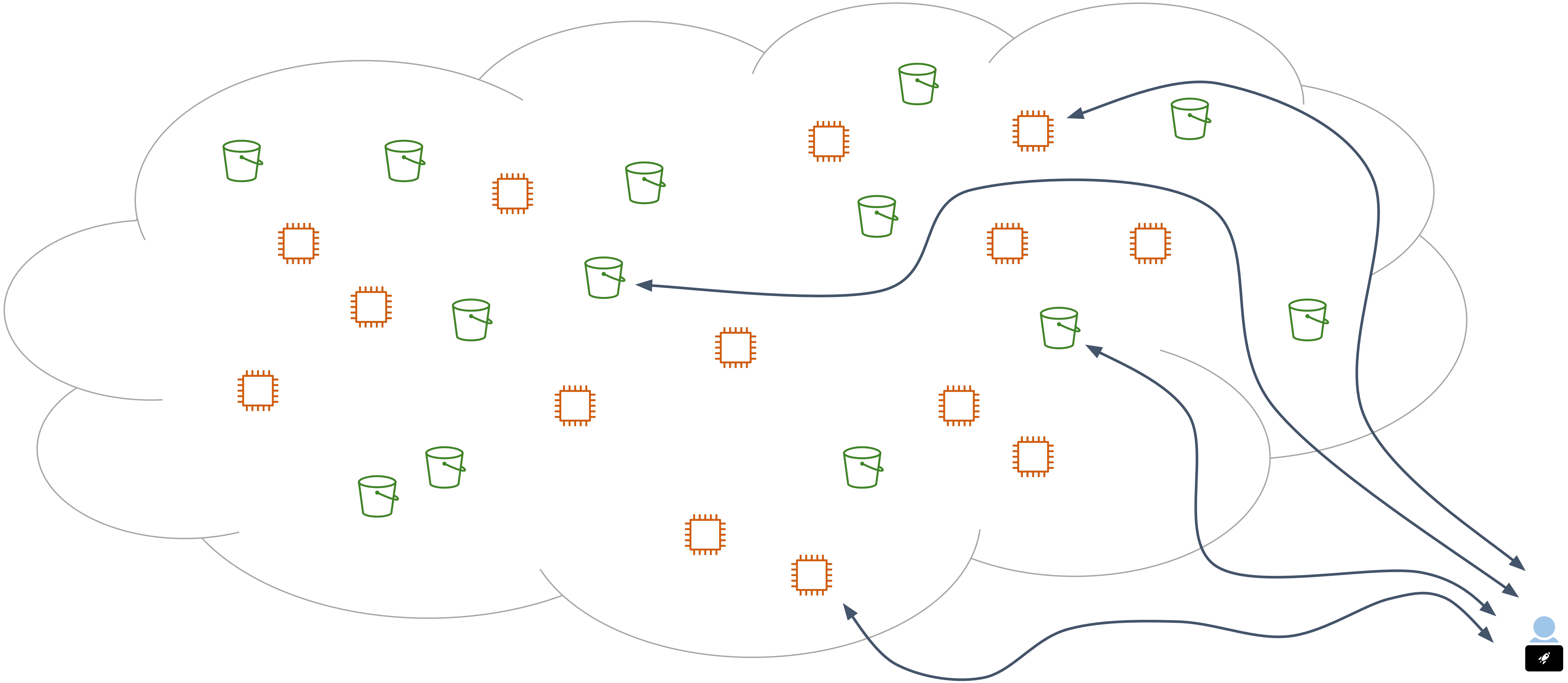
How is the cloud different?



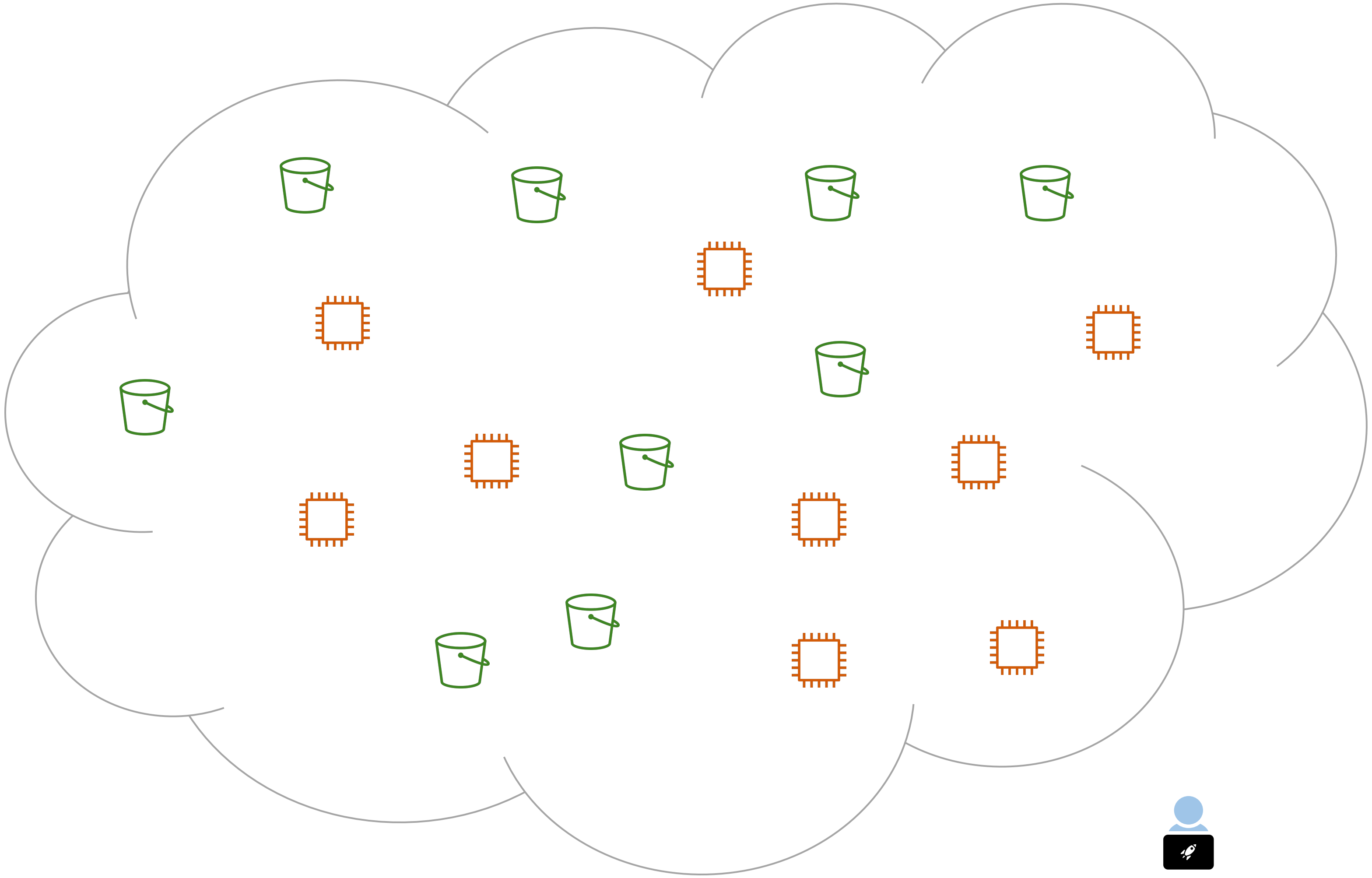
How is the cloud different?



How is the cloud different?

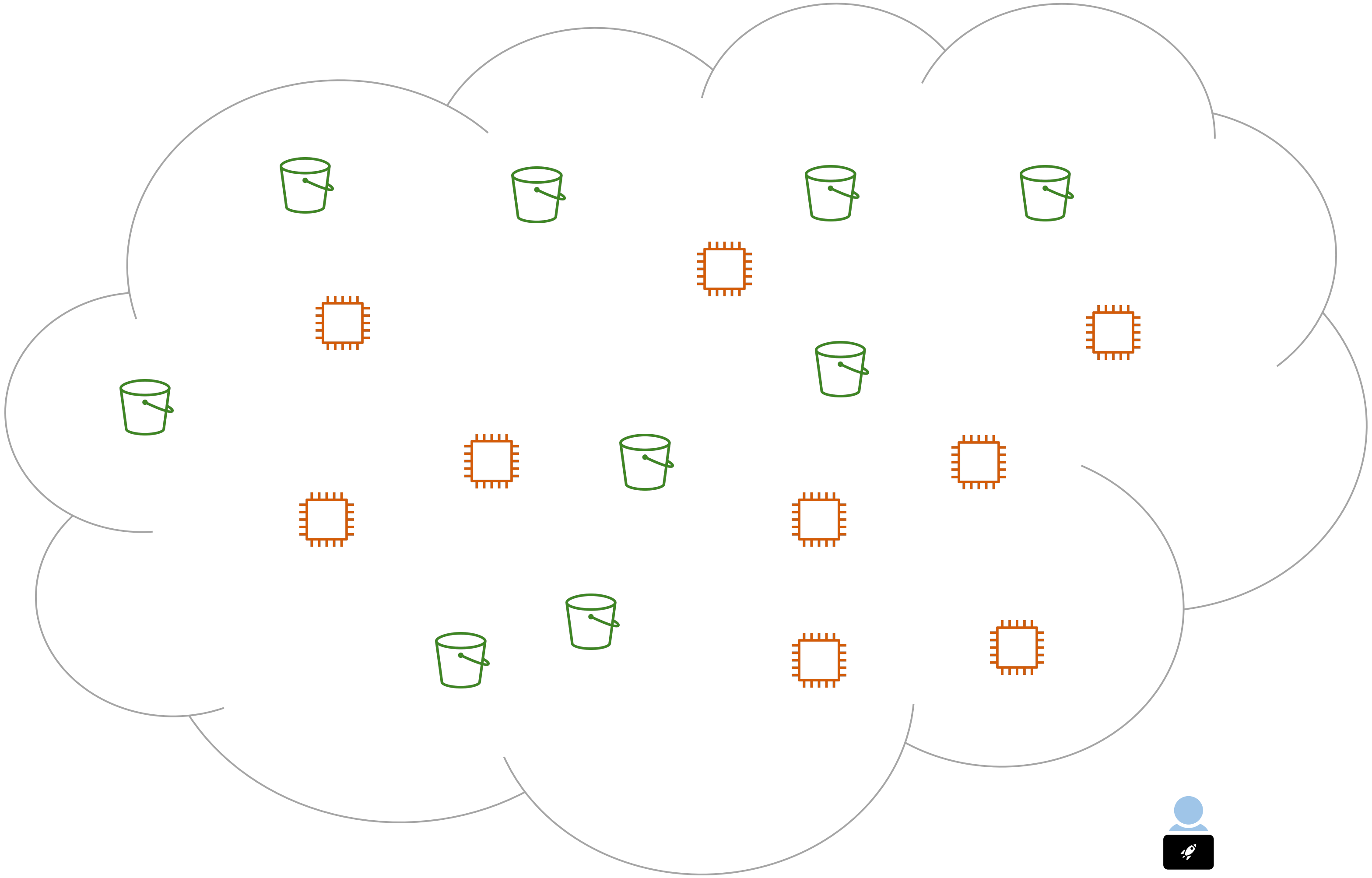


How is the cloud different?



- Compute and data physically far apart (= sloooooow communication)

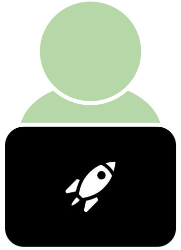
How is the cloud different?



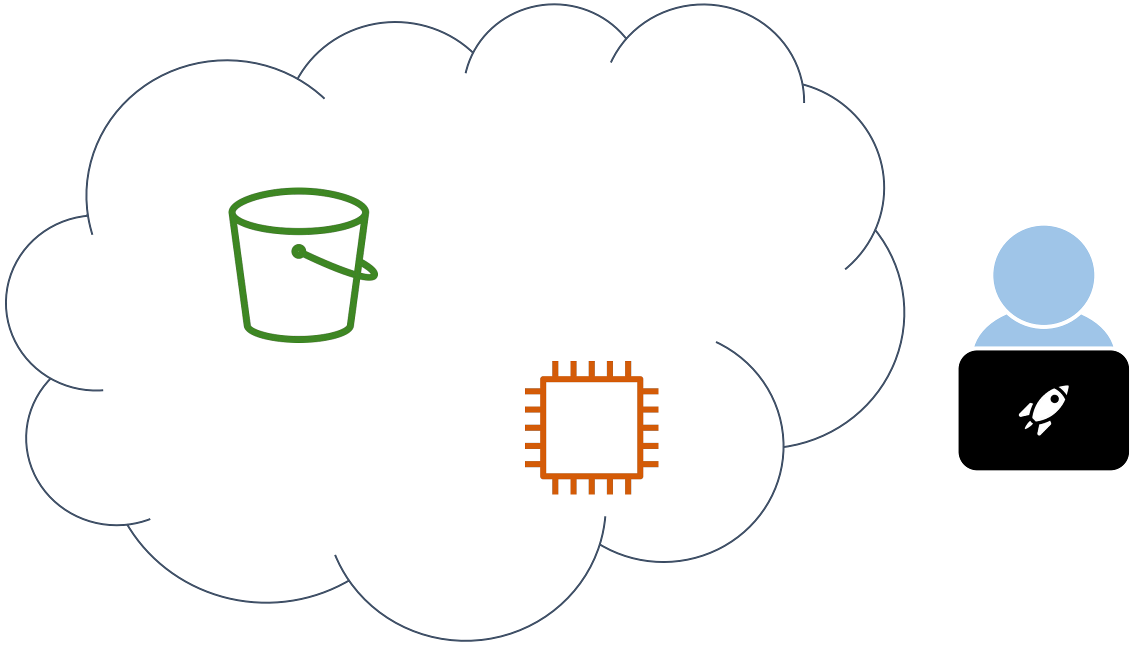
- Compute and data physically far apart (= sloooooow communication)

+ lots of storage
+ lots of compute

How is the cloud different?



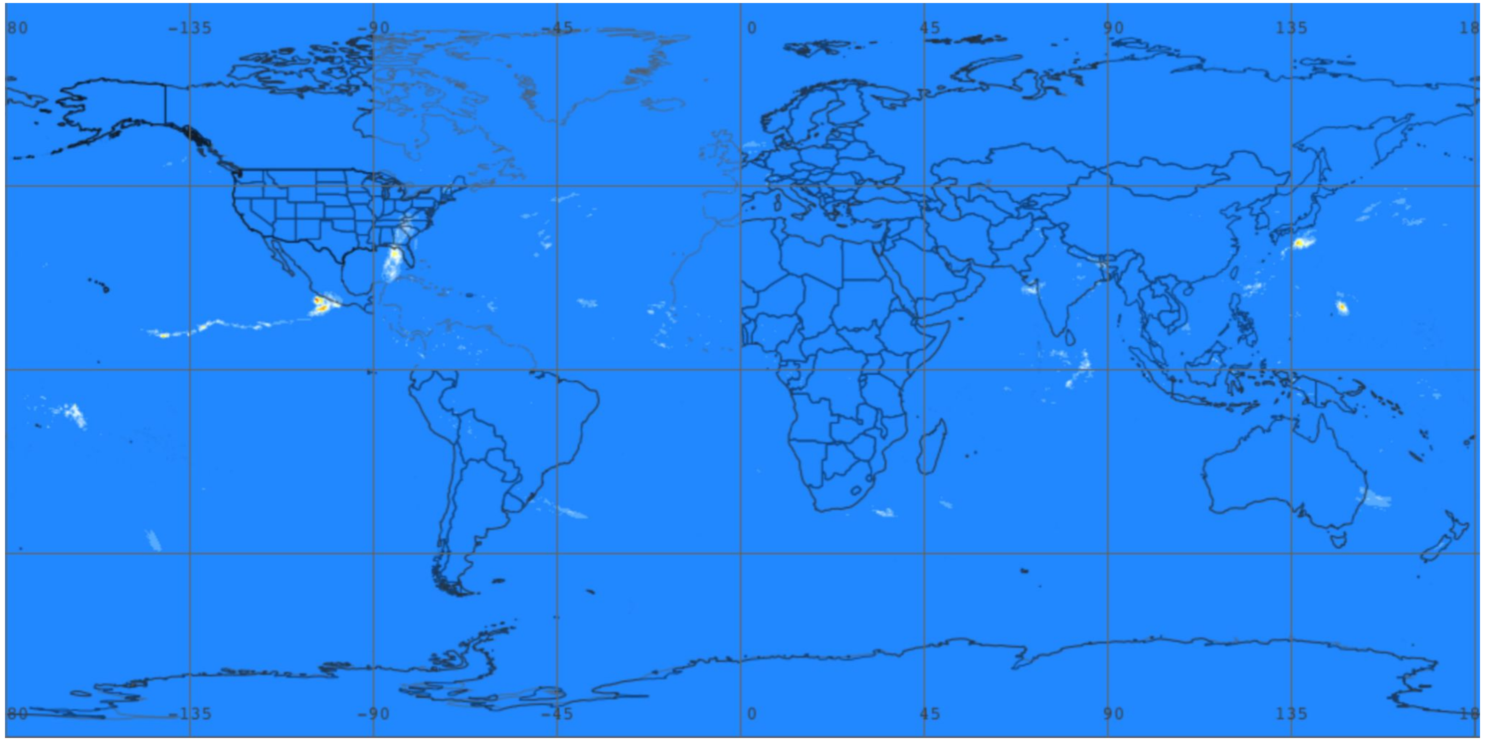
vs.



- + Compute and data physically close together (= FAST Communication!!!)
- - limited storage
- - limited compute

- - Compute and data physically far apart (= slooooooow communication)
- + lots of storage
- + lots of compute

How do you design a data format for the cloud?

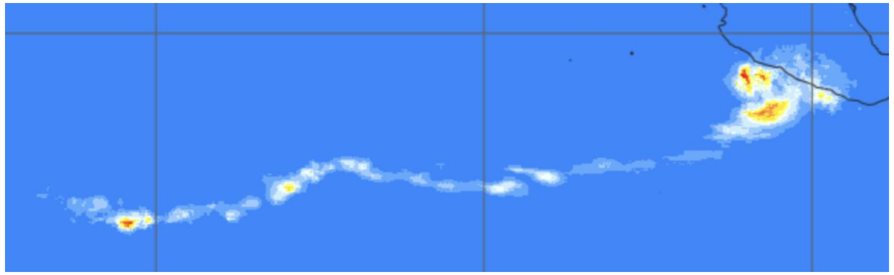


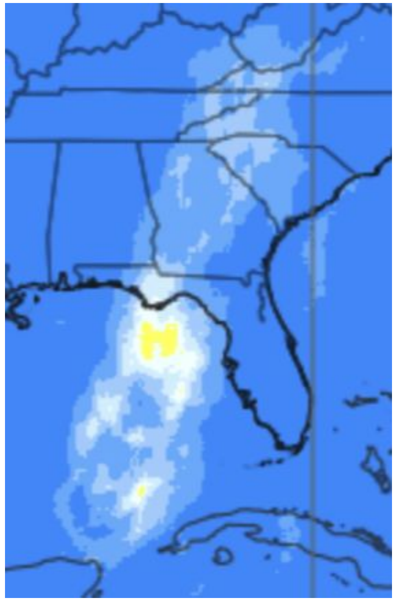
GPM GPM_3IMERGHHL v07 / precipitation

2024-09-26 00:00:00Z - 2024-09-27 01:00:00Z

Optimization 1: Fetch only data you need

How do you design a data format for the cloud?





GPM GPM_3IMERGHHL v07 / precipitation

2024-09-26 00:00:00Z - 2024-09-27 01:00:00Z

Optimization 1: Fetch only data you need

How do you design a data format for the cloud?

Optimization 1: Fetch only data you need

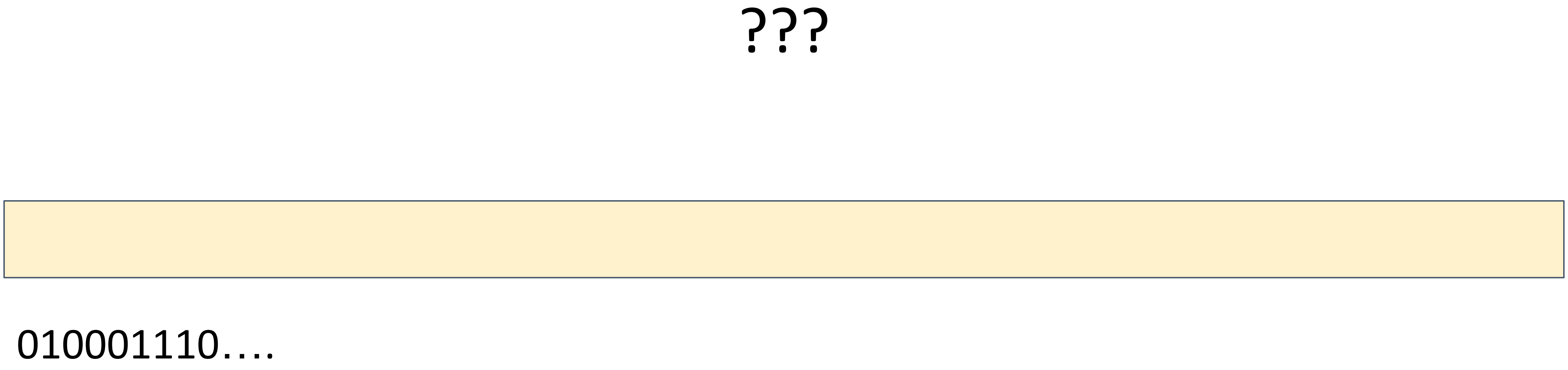
→ **Format Requirement 1**: Use addressable chunks and/or tiling

How do you design a data format for the cloud?

010001110….

Optimization 2: Make metadata easy

How do you design a data format for the cloud?



???

010001110….

Optimization 2: Make metadata easy

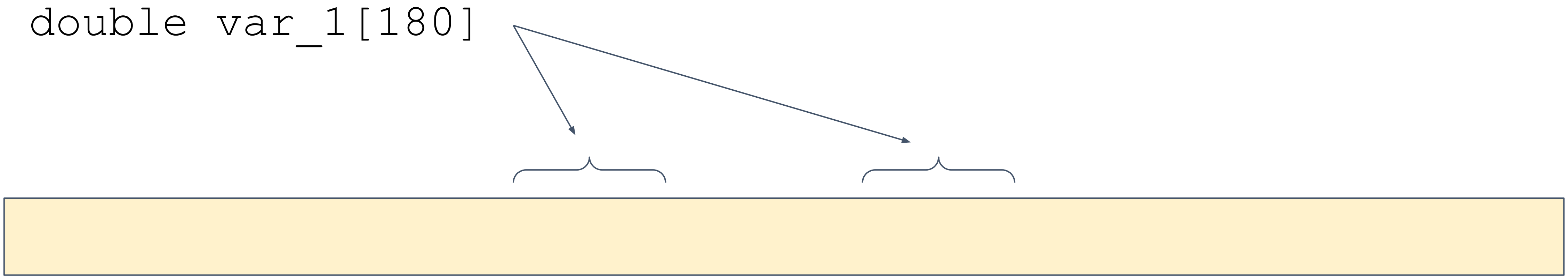
How do you design a data format for the cloud?

```
double var_1[180]
int var_2[360][180]
…
```

010001110….

Optimization 2: Make metadata easy

How do you design a data format for the cloud?



```
double var_1[180]
```

010001110….

Optimization 2: Make metadata easy

Optimization 2: Make metadata easy

→ **Format Requirement 2:** Consolidate metadata for chunks/tiles

How do you design a data format for the cloud?

Format Requirement 1: Use addressable chunks and/or tiling

Format Requirement 2: Consolidate metadata for chunks/tiles

… and now for some actual data formats!

Questions to Ask When Generating Cloud-Optimized Geospatial Data in Any Format

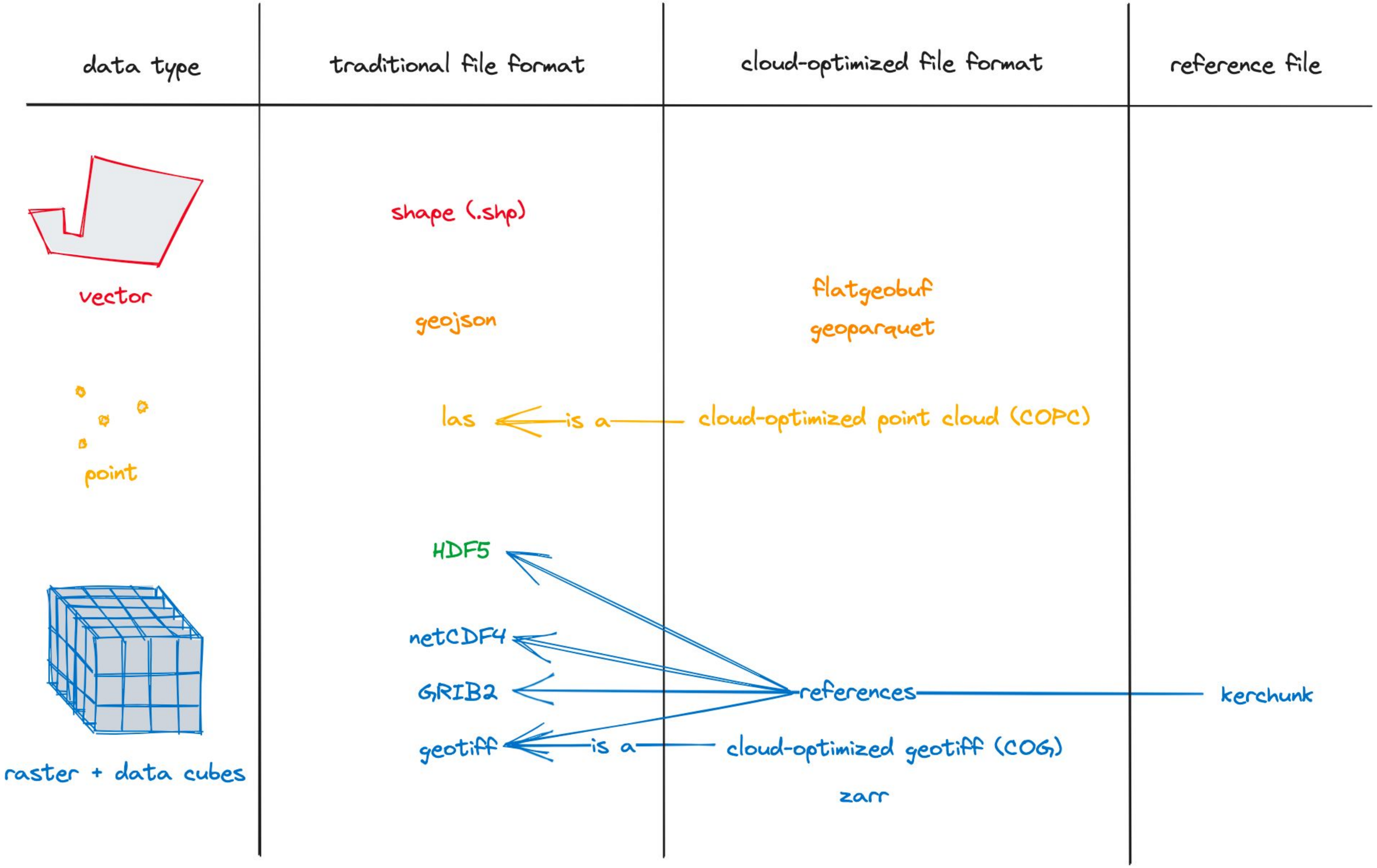
- What variable(s) should be included in the new data format?
- Will you create copies to optimize for different needs?
- What is the intended use case or usage profile? Will this product be used for visualization, analysis, or both?
- What is the expected access method?
- How much of your data is typically rendered or selected at once?

https://guide.cloudnativegeo.org/

What data formats are out there right now?

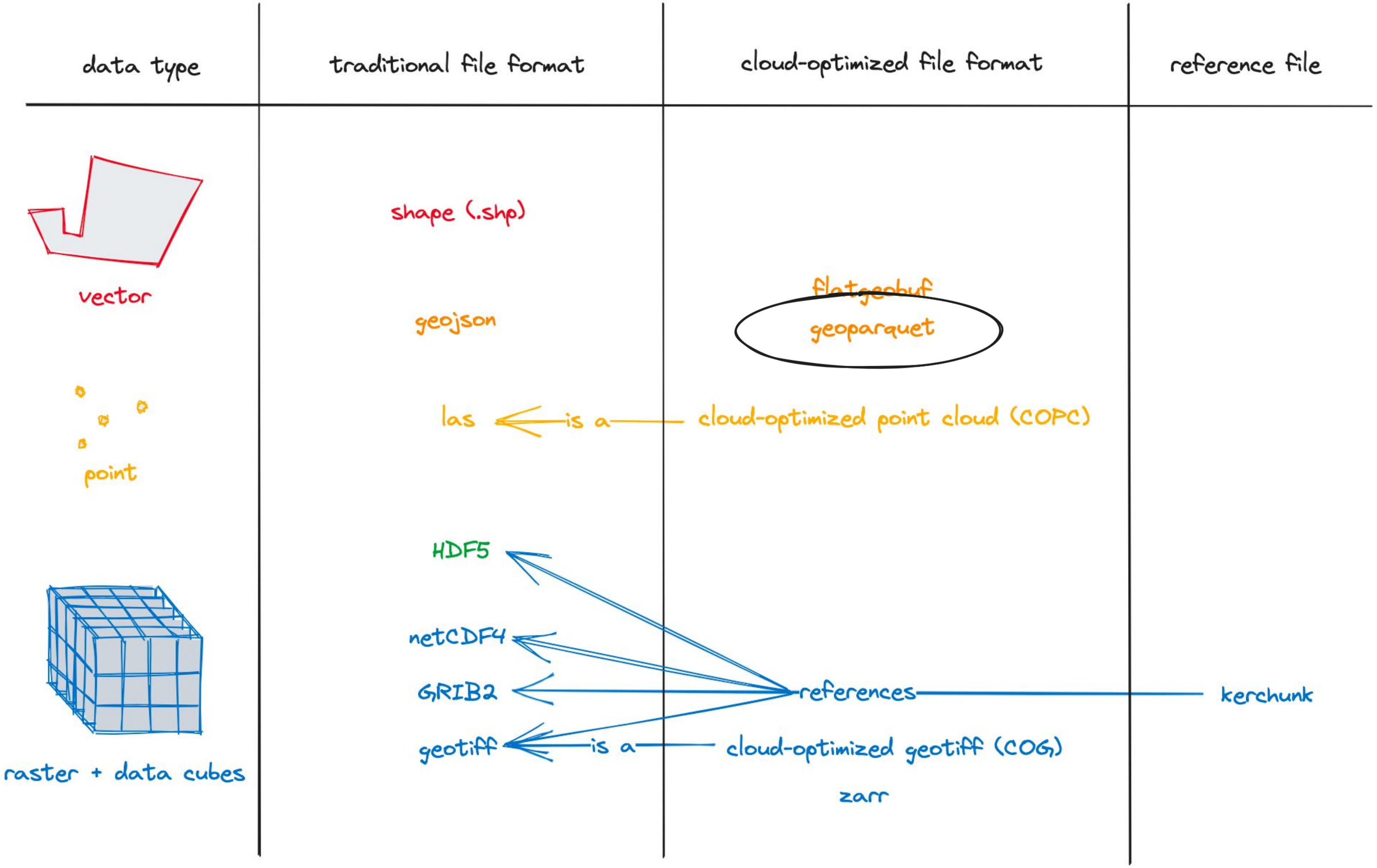
UNDER CONSTRUCTION

What data formats are out there right now?



data type	traditional file format	cloud-optimized file format	reference file
vector	shape (.shp) geojson	flatgeobuf geoparquet	
point	las	cloud-optimized point cloud (COPC)	
raster + data cubes	HDF5 netCDF4 GRIB2 geotiff	references cloud-optimized geotiff (COG) zarr	kerchunk

What data formats are out there right now?



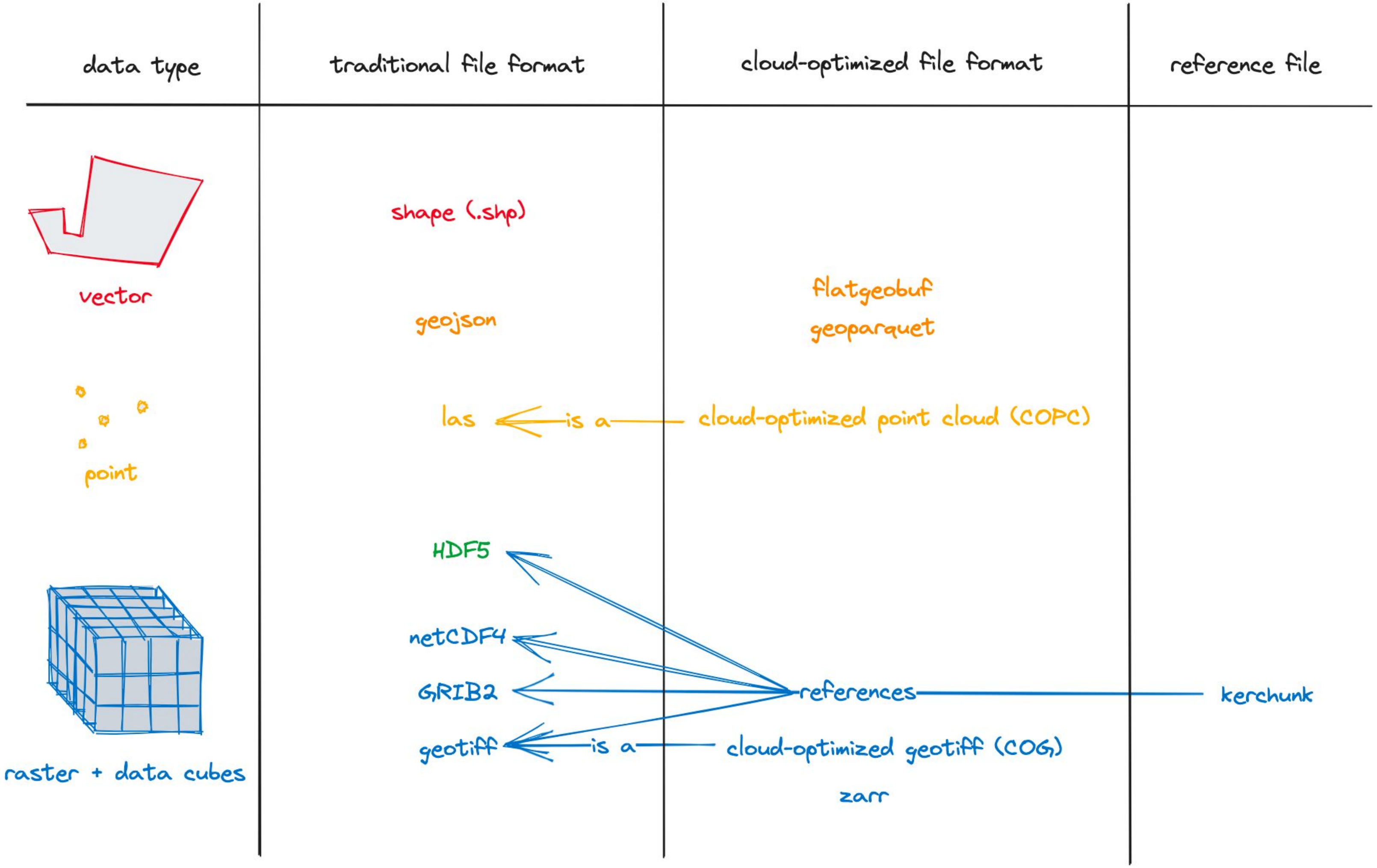
data type	traditional file format	cloud-optimized file format	reference file
vector	shape (.shp) geojson	flatgeobuf geoparquet	
point	las ← is a	cloud-optimized point cloud (COPC)	
raster + data cubes	HDF5 netCDF4 GRIB2 geotiff ← is a	references cloud-optimized geotiff (COG) zarr	kerchunk

GeoParquet

- Apache Parquet with special metadata
- Point, LineString, Polygon, MultiPoint, MultiLineString, MultiPolygon, and GeometryCollection
- Able to specify CRS
- Large files can get cumbersome

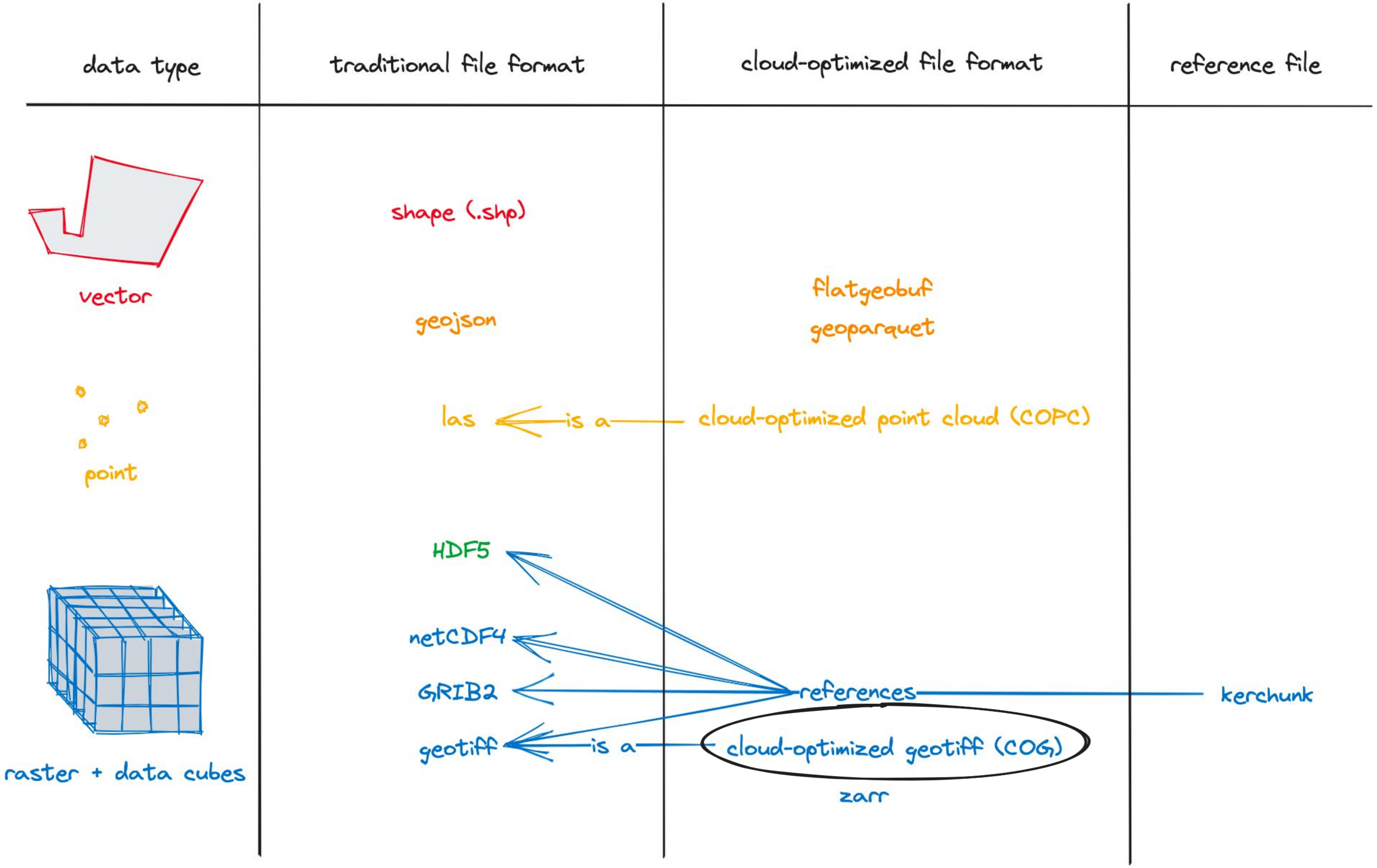
→ Useful for shapes, geojson, and point data

What data formats are out there right now?



data type	traditional file format	cloud-optimized file format	reference file
vector	shape (.shp) geojson	flatgeobuf geoparquet	
point	las ← is a	cloud-optimized point cloud (COPC)	
raster + data cubes	HDF5 netCDF4 GRIB2 geotiff ← is a	references cloud-optimized geotiff (COG) zarr	kerchunk

What data formats are out there right now?

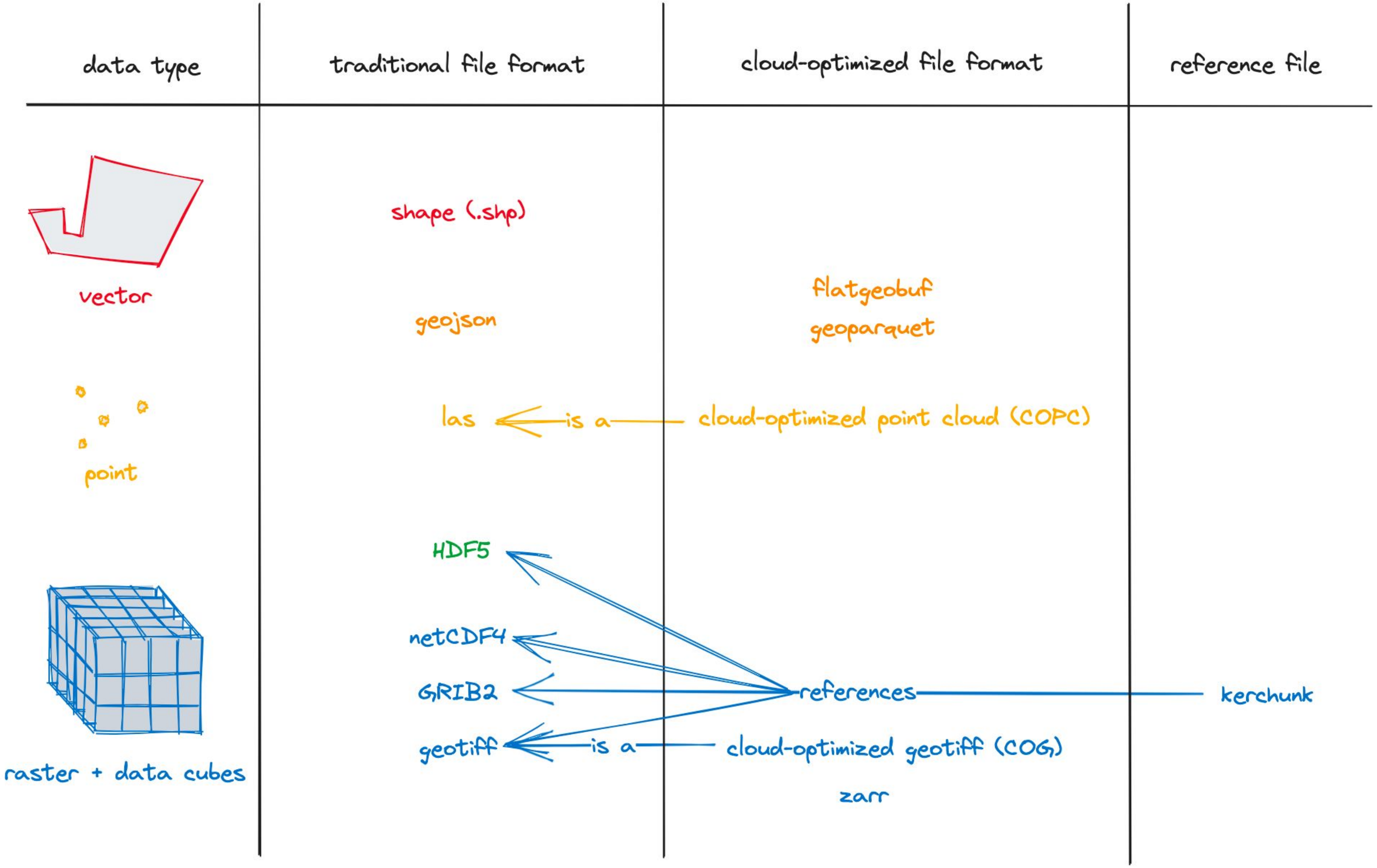


Cloud-Optimized geoTIFF (COG)

- geoTIFF with consolidated metadata
- All COGs are geoTIFFs, so backwards compatible with all your favorite tools

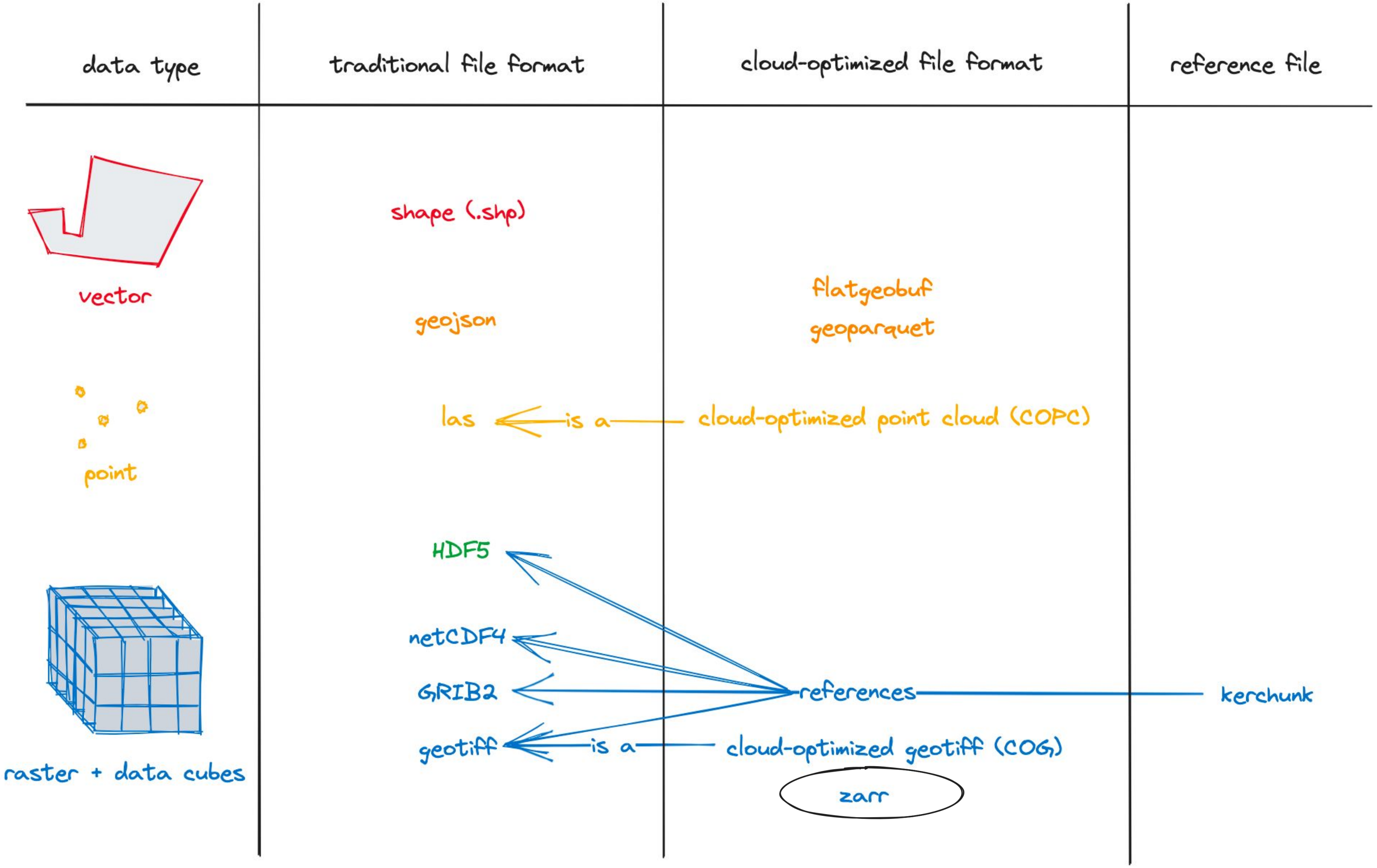
⟶ Good for raster data

What data formats are out there right now?



data type	traditional file format		cloud-optimized file format	reference file
	shape (.shp)			
vector	geojson		flatgeobuf geoparquet	
point	las	← is a —	cloud-optimized point cloud (COPC)	
raster + data cubes	HDF5			
	netCDF4			
	GRIB2		references	kerchunk
	geotiff	← is a —	cloud-optimized geotiff (COG)	
			zarr	

What data formats are out there right now?



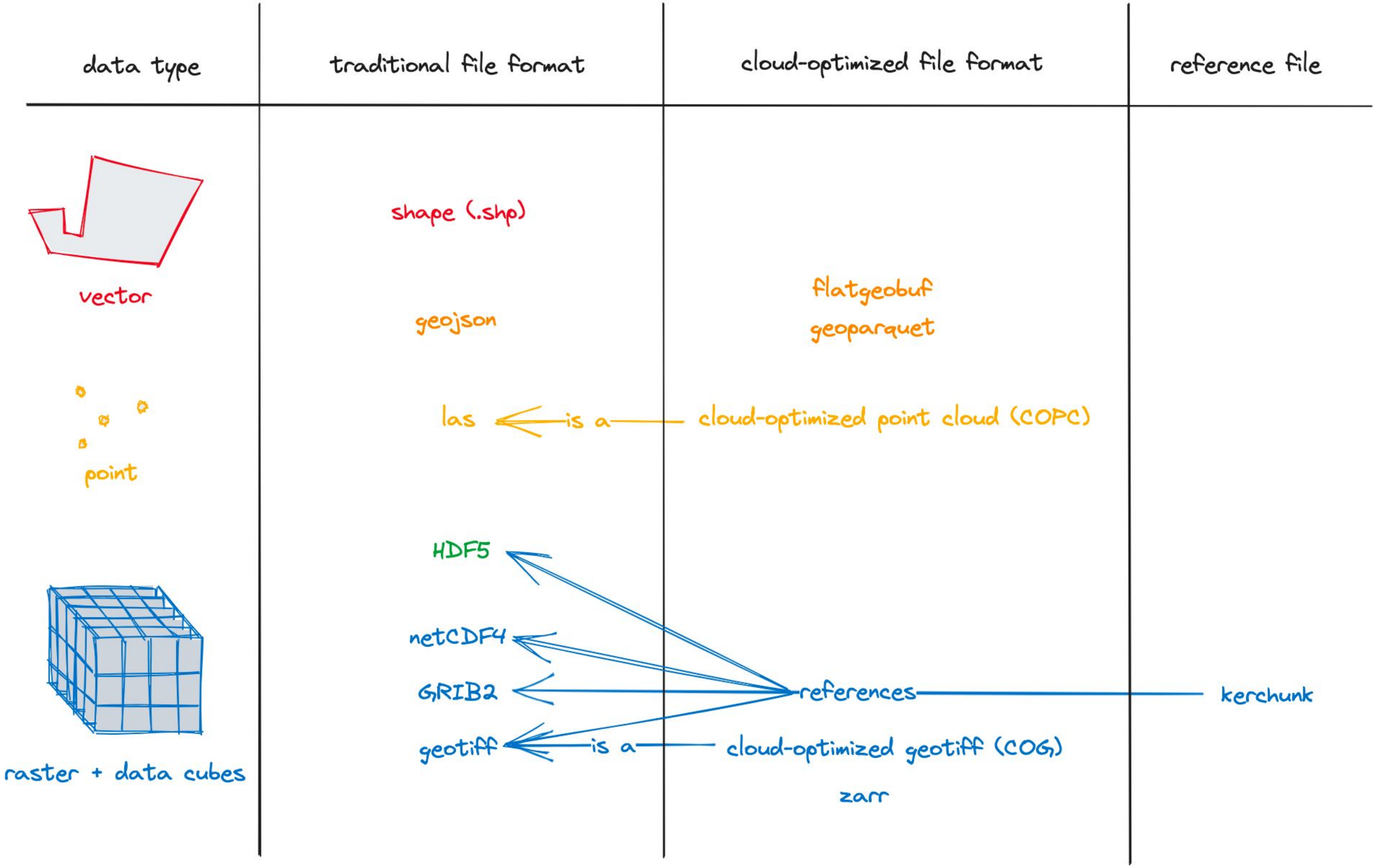
data type	traditional file format	cloud-optimized file format	reference file
vector	shape (.shp)		
	geojson	flatgeobuf geoparquet	
point	las ← is a —	cloud-optimized point cloud (COPC)	
raster + data cubes	HDF5 netCDF4 GRIB2 geotiff ← is a —	references cloud-optimized geotiff (COG) zarr	kerchunk

Zarr

- Basically the same data model as HDF: arrays + metadata
- Really shines with very, very (VERY!) large arrays
- Possible to update, but tricky to manage with multiple reader/writers

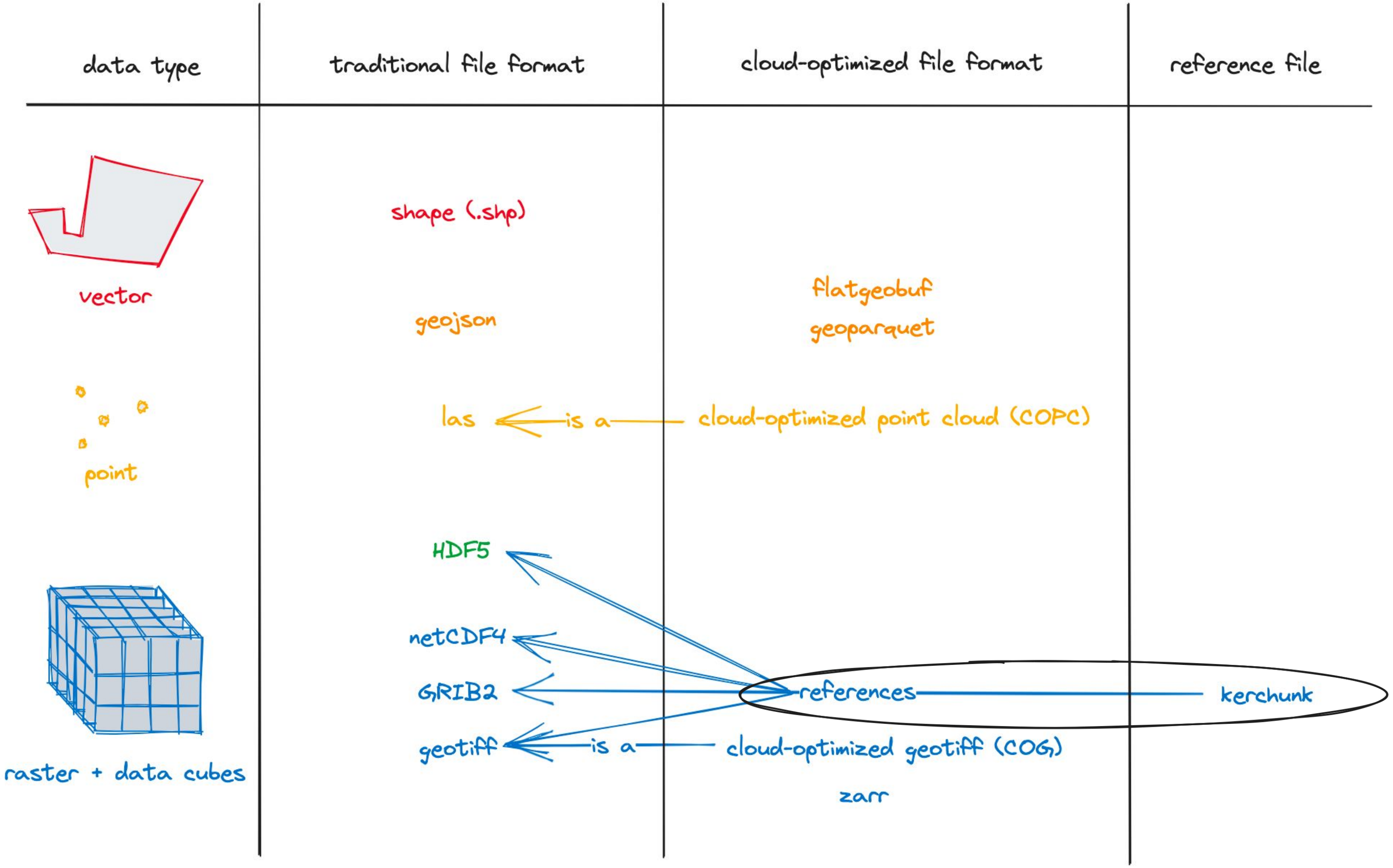
→ Good for multi-dimensional arrays

What data formats are out there right now?



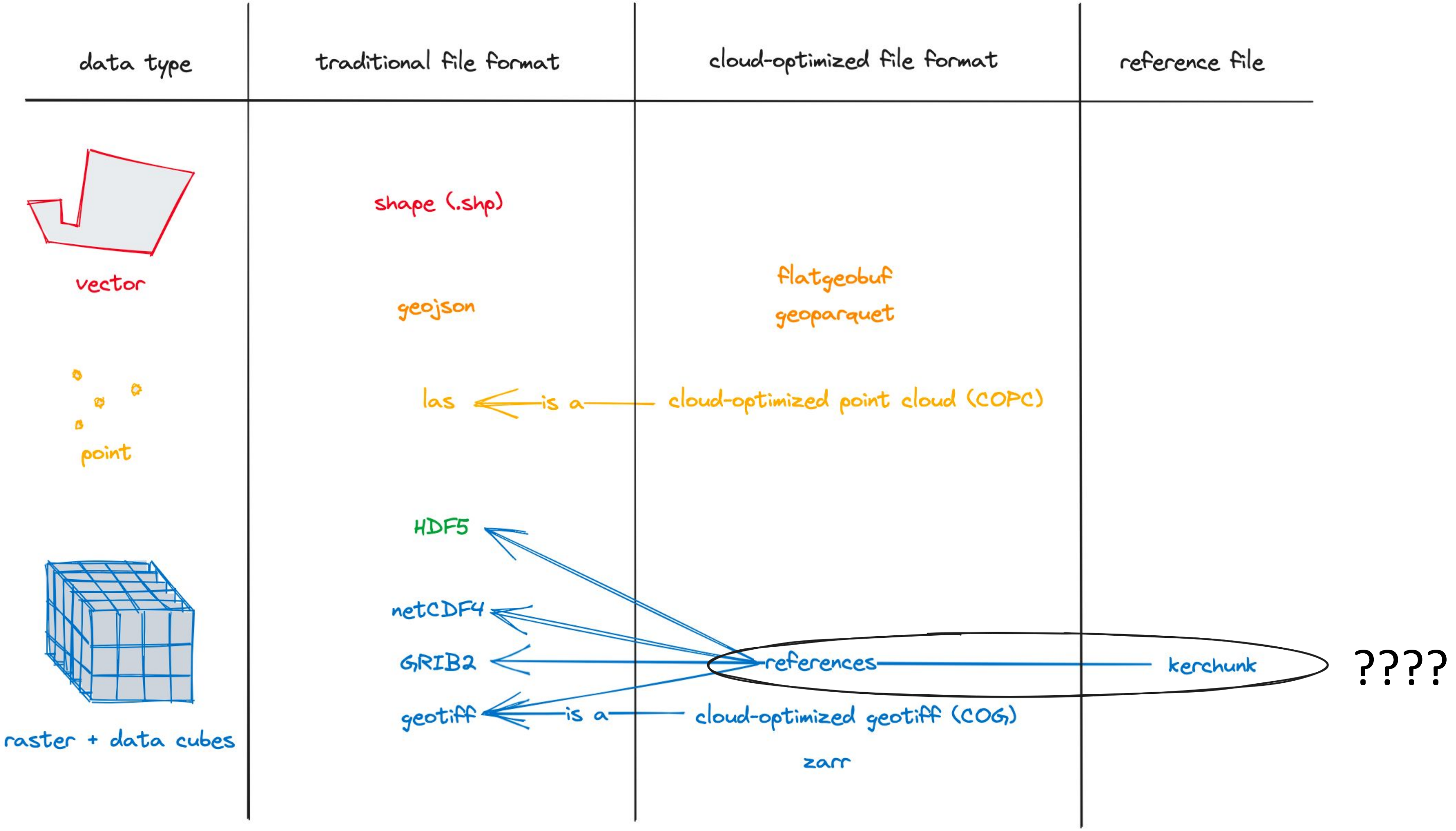
data type	traditional file format	cloud-optimized file format	reference file
vector	shape (.shp) geojson	flatgeobuf geoparquet	
point	las ←is a—	cloud-optimized point cloud (COPC)	
raster + data cubes	HDF5 netCDF4 GRIB2 geotiff ←is a—	←references— cloud-optimized geotiff (COG) zarr	kerchunk

What data formats are out there right now?



data type	traditional file format		cloud-optimized file format	reference file
vector	shape (.shp)			
	geojson		flatgeobuf geoparquet	
point	las	is a	cloud-optimized point cloud (COPC)	
raster + data cubes	HDF5 netCDF4 GRIB2		references	kerchunk
	geotiff	is a	cloud-optimized geotiff (COG)	
			zarr	

What data formats are out there right now?



data type	traditional file format	cloud-optimized file format	reference file
vector	shape (.shp) geojson	flatgeobuf geoparquet	
point	las	cloud-optimized point cloud (COPC)	
raster + data cubes	HDF5 netCDF4 GRIB2 geotiff	cloud-optimized geotiff (COG) zarr	kerchunk

References & Kerchunk

Format Rule 1: Use addressable chunks and/or tiling

Format Rule 2: Consolidate metadata for chunks/tiles

References & Kerchunk

	Chunks/Tiling?	Consolidated metadata?
HDF5		
netCDF4		
GRIB2		
geoTIFF		

References & Kerchunk

	Chunks/Tiling?	Consolidated metadata?
HDF5	✅	
netCDF4		
GRIB2		
geoTIFF		

References & Kerchunk

	Chunks/Tiling?	Consolidated metadata?
HDF5	✅	
netCDF4	✅	
GRIB2		
geoTIFF		

References & Kerchunk

	Chunks/Tiling?	Consolidated metadata?
HDF5	✅	
netCDF4	✅	
GRIB2	✅	
geoTIFF		

References & Kerchunk

	Chunks/Tiling?	Consolidated metadata?
HDF5	✅	
netCDF4	✅	
GRIB2	✅	
geoTIFF	✅	

References & Kerchunk

	Chunks/Tiling?	Consolidated metadata?
HDF5	✅	❌
netCDF4	✅	
GRIB2	✅	
geoTIFF	✅	

References & Kerchunk

	Chunks/Tiling?	Consolidated metadata?
HDF5	✅	❌
netCDF4	✅	❌
GRIB2	✅	
geoTIFF	✅	

References & Kerchunk

	Chunks/Tiling?	Consolidated metadata?
HDF5	✅	❌
netCDF4	✅	❌
GRIB2	✅	❌
geoTIFF	✅	

References & Kerchunk

	Chunks/Tiling?	Consolidated metadata?
HDF5	✅	❌
netCDF4	✅	❌
GRIB2	✅	❌
geoTIFF	✅	❌

References & Kerchunk

	Chunks/Tiling?	Consolidated metadata?
HDF5	✅	❌
netCDF4	✅	❌
GRIB2	✅	❌
geoTIFF	✅	❌*

* Original geoTIFF

References & Kerchunk

	Chunks/Tiling?	Consolidated metadata?
HDF5	✅	❌
netCDF4	✅	❌
GRIB2	✅	❌
geoTIFF	✅	❌

References & Kerchunk

	Chunks/Tiling?	Consolidated metadata?
HDF5	✅	❌
netCDF4	✅	❌
GRIB2	✅	❌
geoTIFF	✅	❌

References & Kerchunk

	Chunks/Tiling?	Consolidated metadata?
HDF5	✅	❌
netCDF4	✅	❌
GRIB2	✅	❌
geoTIFF	✅	❌

Solution: Create a separate metadata file!

References & Kerchunk

- Make your existing data files cloud-optimized - avoid making copies!
- Compatible with zarr libraries
- Limited by existing chunking in files

→ Good for making your existing raster/data cube data perform better in the cloud

Watch this space!

UNDER
CONSTRUCTION

Thanks!!!

- Wikimedia Commons: laptop components, server racks
- Giovanni: precipitation map
- Amazon: AWS bucket and EC2 icons
- Cloud-Optimized Geospatial Formats Guide: Geospatial File Format Table
- Fellow engineers at the GES DISC who have also worked hard on data in the cloud, particularly Hailiang Zhang and Dieu My Nguygen

Questions?