

# AI in The Sky =? Pie in The Sky

Kees van Berkel

MPSoC 2026  
Victoria BC, Canada

Consultant at Snap Inc.  
Emeritus professor TU Eindhoven



# Elon Musk's view on Starcloud's Orbital AI Vision

“The basic math is that launching **a million tons per year** of satellites generating **100 kW of compute power per ton** would add **100 gigawatts of AI compute** capacity annually, with no ongoing operational or maintenance needs.”

“My estimate is that within 2 to 3 years, the **lowest cost** way to generate AI compute will be in space.”

“Ultimately, there is a path to launching **1 TW/year** from Earth.”

Verbatim from [2025, SpaceX].

SpaceX seeks FCC permission to put up to **a million data center satellites** in orbit, [2026, SpaceX, FCC application].



# A 1 GW AI Data Center, terrestrial

| Capex                     | Cost [B\$]   | fraction [%] |
|---------------------------|--------------|--------------|
| Compute hardware          | 12–20        | 35–40        |
| Electrical infrastructure | 14–18        | 40–45        |
| Cooling systems           | 5–10         | 15–20        |
| Land + construction       | 7–12         | 10–15        |
| Networking                | 1–2          | 3–5          |
| Water + utilities         | 0.5–1        | 1–3          |
| <b>Total</b>              | <b>35–60</b> | <b>100</b>   |

| Opex         | Annual Cost [B\$] | fraction [%] |
|--------------|-------------------|--------------|
| Energy       | 0.6               | 65–70        |
| Labor        | 0.1–0.2           | 10–20        |
| Maintenance  | 0.1               | 10–12        |
| Taxes + land | 0.05–0.1          | 5–8          |
| <b>Total</b> | <b>0.9</b>        | <b>100</b>   |

Sources: Bernstein Research, NVIDIA, Epoch AI

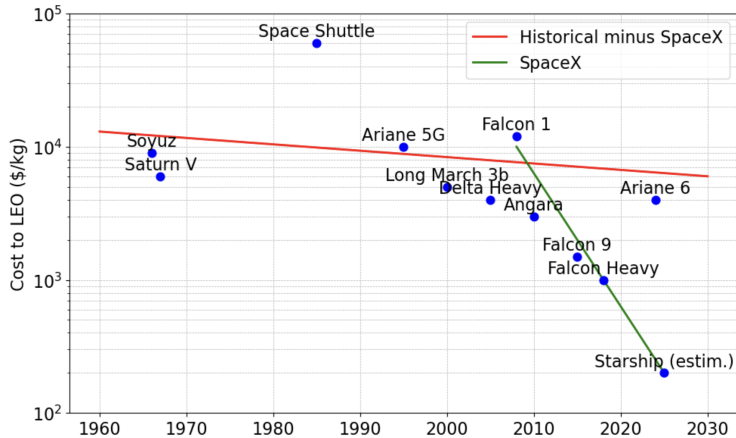


OpenAI, Ann Arbor, Michigan, 1GW, under construction



Meta, Lebanon, Indiana, 1GW, under construction

# Cost of Launch to Low Earth Orbit



Source: [Marspedia]



Starship V2

- 1GW of AI compute
- = 10,000 ton of satellites (100 kW of compute power per ton)
- = 50 Starship V3 payloads (200 ton payload)
- = 2B\$ launch costs (200\$ per kg)



# Starlink Satellites, Solar Power

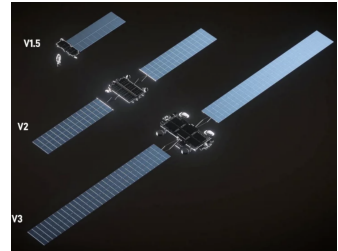
- ▶ >10k satellites in Low Earth Orbit  $\approx 600\text{km}$ .
- ▶ 12-16 Falcon 9 launches per month in 2025.
- ▶ 20-29 Starlink V2 Mini satellites per launch.

Solar irradiance in LEO orbit:  $\approx 1300 \text{ W/m}^2$   
Solar panel efficiency (guess)  $\approx 30 \%$   
Solar power per  $\text{m}^2$  (used below)  $\approx 400 \text{ W/m}^2$

| Starlink | Mass [kg] | Area [ $\text{m}^2$ ] | Power [kW] |
|----------|-----------|-----------------------|------------|
| V1       | 300       | 25                    | 10         |
| V2 mini  | 800       | 100                   | 40         |
| V3       | 2000      | 250                   | 100        |

No atmosphere, always normal to the sun.

- Starlink orbits: day-night cycle.
- **Sun-synchronous (polar) orbits**:  $400 \text{ W/m}^2$ , all day.



Starlink V1.5 - V3 satellites.



Starlink V2-mini, launch cfg.  
(1 tray = 1 satellite)

# Cooling

$P = 40\text{kW}$  of cooling in space: not by convection or conduction, ...  
but only by **radiative cooling**, governed by the *Stefan–Boltzmann law*:

$$P = \sigma \cdot A \cdot T^4, \quad \text{with } \sigma \approx 5.67 \cdot 10^{-8} \text{ W} \cdot \text{m}^{-2} \text{K}^{-4}$$

Musk on X, 2026-Jan-20: “GPU designed to operate at  $T=370\text{K}$ ”.

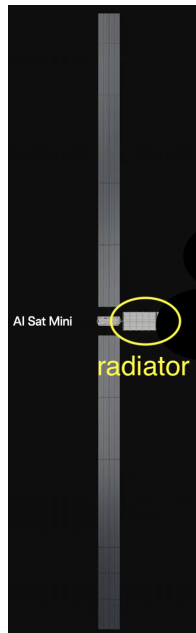
- $\Rightarrow A = 2 \times 20\text{m}^2$ , Starlink V2 mini, 2-sided radiator.
- ... ignoring absorption, heat transport, emissivity, ...

Starlink radiator  $A$  looks  $3 \times$  **too small** vs  $100\text{m}^2$  of solar panels:

- ... solar efficiency  $< 30\%$ ? charging a battery? radio power?
- ... “patterned cellular grid fin structure”? [2020, SpaceX patent]

Scaling  $> 100\text{kW}$ :

- Spread GPUs evenly across area larger radiator (chassis for compute).
- Active cooling (power hungry!) looks avoidable.



# Radiation Tolerance

- ▶ Radiation (energetic protons, alpha particles) causes **Total Ionizing Dose effects** (TID), causing device degradation and **Single Event Effects** (SEE) causing (multiple) bit errors.
- ▶ Starlink uses radiation hardened components, incl. STM32V8 MCU (STM, 18nm SOI) and Versal (AMD/Xilinx, 7nm CMOS).
- ▶ Radiation hardening is costly: multiple CMOS nodes behind,  $2\times$  area,  $0.8\times$  clock, ... not affordable.

[2025, Google] tested its **V6 TPU** in a **67 MeV proton beam**  $\approx$  sun-synchronous LEO:

- ▶ Using **shielding** (10 mm Al equivalent)  $\Rightarrow$  dose  $\approx$  **150 rad/year**.
- ▶ Over 5 year lifetime a dose of **750 rad** looks acceptable.  
(HBM-specific stress tests: irregularities after TID  $> 2$  krad,)
- ▶ SEEs must be at levels manageable through **memory error correcting codes** (ECC).

(Google's version of orbital AI is named *Suncatcher*.)

# (Unintended) Electromagnetic Radiation

100 GW  $\times$  5 (5y lifespan)

$\Rightarrow$  500 GW infrared radiation:

$\Rightarrow$  headache for infrared astronomy.

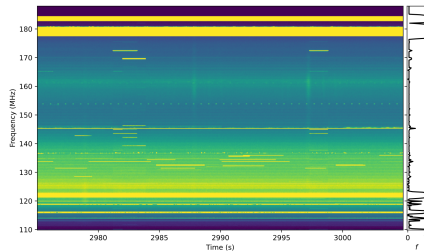
Starlink: *unintended electromagnetic radiation*,

$\Rightarrow$  headache for radio astronomy [2023, LOFAR].

$\times 100$  (?) for in-orbit AI.



Starlink sky, 2021.



LOFAR: 3 Starlink satellites, 2023.

# Reentry

A Starlink satellite has a life time of  $\approx 5$  years.

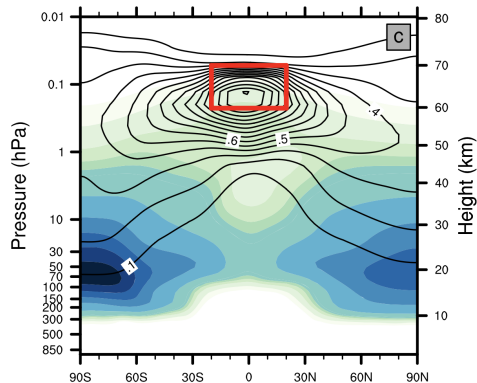
Goal: de-orbit leading to 100% vaporization

$\Rightarrow$   $Al_2O_3$  aerosol (nm- $\mu$ m particulates).

[2024 NOAA (US Gov)] simulations:

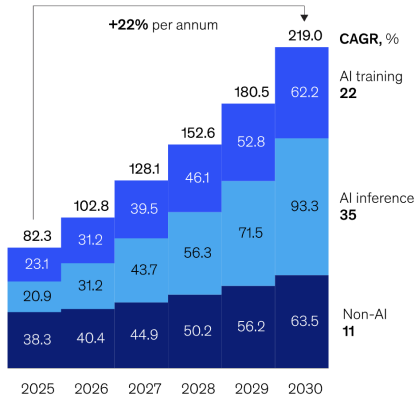
- ▶ Assume a mass flux of  $10^4$  ton/yr.
- ▶  $Al_2O_3$  aerosols accumulate poleward of  $30^\circ$  N/S, between 10 - 30 km.
- ▶ Mesospheric heating rates lead to  $1.5^\circ$ K temperature anomalies.
- ▶ These anomalies lead to a weaker springtime ozone hole.

Musk plans  $10^6 - 10^7$  ton/yr. Plus: orbital AI by Amazon, Google, China, ...

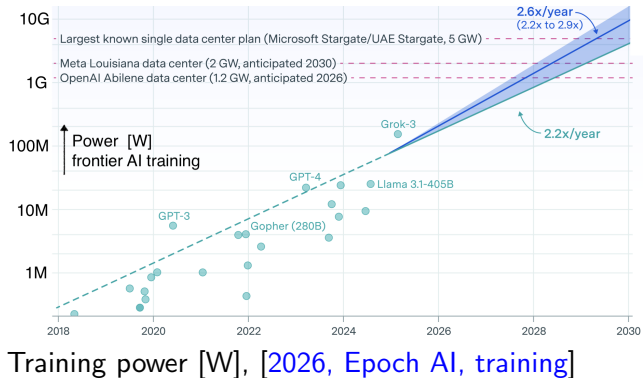


$Al_2O_3$  aerosol, simulation,  
controlled reentry @ equator.

# AI Compute: Workloads



Global [GW], [2026, McKinsey]



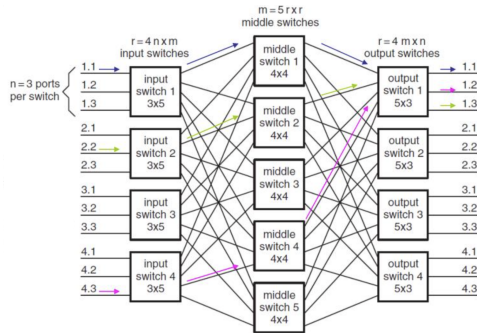
- In 2030: both AI training and AI inference are below Musk's 100GW/yr.
- Epoch AI: 2.1x/y vs McKinsey 1.22x/y  $\Rightarrow$  ever larger, but fewer models/y.

# AI compute: 1GW Satellite Constellations and Network Topologies

## Terrestrial 1GW

≈ 8000 racks × 120kW (e.g. Blackwell)

≈ 600k GPUs ≈ 5600 EFLOP/s.

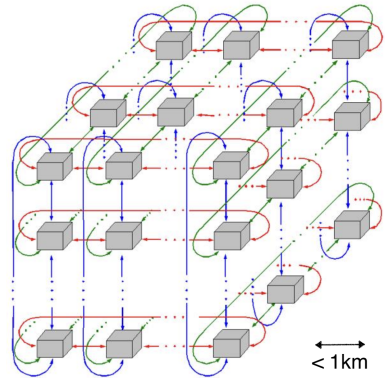


Topology: folded **Clos network** ≈ fat tree.

- optical (Ethernet) links.

## In orbit (speculated constellation) 1GW

≈ 8000 satellites × 100kW (Starlink V3)



Topology: **3D torus**,  $20 \times 20 \times 20$ .

- Laser Inter-Satellite Links (LISL).

# AI Compute: Performance Model

$$T = T_{latency} + T_{bandwidth} + T_{compute}$$

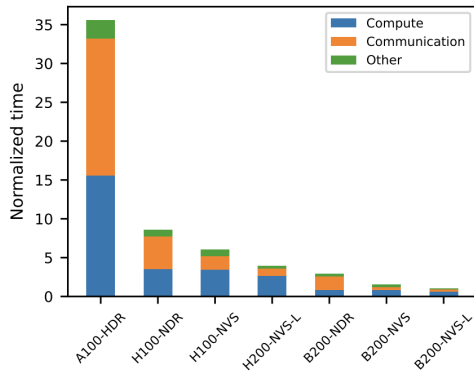
$$= 2L \log_2(N) + \frac{2K}{N \times B}(N-1) + \gamma \frac{1}{N}$$

|            |               |       |                 |                                         |
|------------|---------------|-------|-----------------|-----------------------------------------|
| training:  | $T_{latency}$ | $\ll$ | $T_{bandwidth}$ | $K$ large                               |
| inference: | $T_{latency}$ | $>$   | $T_{bandwidth}$ | $K$ small                               |
| measured:  | $T_{latency}$ | $+$   | $T_{bandwidth}$ | $= \kappa T,$ with $0.2 < \kappa < 0.6$ |

$N$ : number of nodes,  $K$ : data volume,  
 $L$ : network latency,  $B$ : network bandwidth,  
 $\kappa$ : fraction of  $T$  spent on communication.

Based on [2024, IMEC]

Performance model for training+inference.





# AI Compute: Network Fabrics for 1GW

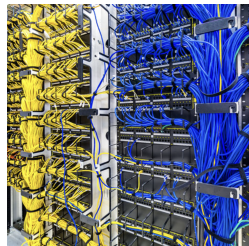
**Bisection bandwidth  $B$ :** bandwidth from/to half the nodes to/from the other half.

| network topology           | unit    | terrest.<br>Clos*                 | orbital<br>2D torus          | orbital<br>3D torus           |
|----------------------------|---------|-----------------------------------|------------------------------|-------------------------------|
| $B$ , bisection bw.        |         | $\frac{1}{2} N \cdot W \cdot B_l$ | $2N^{1/2} \cdot W \cdot B_l$ | $2N^{2/3} \cdot W \cdot B_l$  |
| $W$ , width ToR sw.        | link    | 72**                              | 1                            | 1                             |
| $B_l$ , link bandwidth     | GB/s    | 100                               | 100***                       | 100***                        |
| $B$ , bisection bw.        | TB/s    | 28800                             | 18                           | 80                            |
| $T_{bandwidth}$ , relative |         | 1                                 | 1600                         | 360                           |
| $L$ , network latency      |         | $\log(N) \cdot D$                 | $N^{1/2} \cdot D$            | $\frac{3}{2} N^{1/3} \cdot D$ |
| $D$ , link latency         | $\mu s$ | 1                                 | 20                           | 20                            |
| $L$ , network latency      | $\mu s$ | 4                                 | 1800                         | 600                           |
| $T_{latency}$ , relative   |         | 1                                 | 450                          | 150                           |

\*Clos oversubscription 2:1.

\*\* NVIDIA Blackwell.

\*\*\*8x Starlink V2.



# AI compute: Inference

- ▶ 8 satellites needed for 1T parameters;  $64\times$  batching (for weight reuse), 8 replicas (for load balancing), low latency (for interactive use).
- ▶ 1 satellite inference should be “fairly straightforward” to achieve: larger satellites (more kW) and dedicate inference processors (more flops/W).

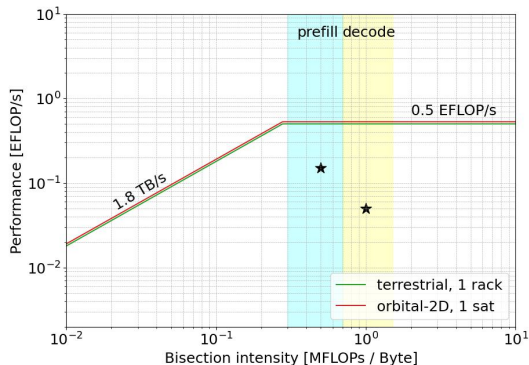
Bisection (arithmetic) intensity:

$$= \frac{\text{total FLOPs}}{\text{data volume that crosses the bisection}}$$

[MFLOPs/Byte] (cf. arithmetic intensity).

Inference

$$= \text{prefill (input phase)} \\ + \text{decode (output phase, memory bound)}$$



# AI compute: Training

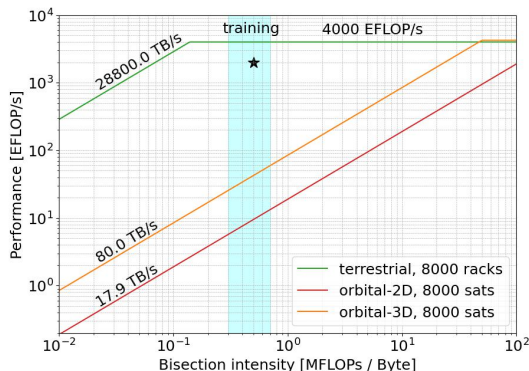
Bisection (arithmetic) intensity:

$$= \frac{\text{total FLOPs}}{\text{data volume that crosses the bisection}}$$

[MFLOPs/Byte] (cf. arithmetic intensity).

Bisection data

= local gradients during back propagation.  
(12B: weights + gradients + Adam moments)



Training in orbit appears network bound by about 2 orders of magnitude.

Moreover:

- In the popular *bucket algorithm* for *torus all-reduce*: data crosses bisections  $3\times$ .
- The impact of  $L \times 150$  is not included in roofline models.

# AI compute: Training; Can it be Fixed?

## ► Fewer and larger satellites?

- E.g.  $8\times$  more compute per satellite (more kW/sat and/or more FLOPs/W).
- $20\times 20\times 20 \rightarrow 10\times 10\times 10$  satellites: bisection bandwidth  $B \times \frac{1}{4}!$

## ► Free-space optics? [2025, Google]

- $B \times 10$  from multi wavelength light beam? Another  $B \times 10$  from multi-beam?
- *Even if* successful: the next challenge: network latency  $L$ .

## ► A “Clos satellite constellation”? (mimicking the Clos network topology)

- 8000 compute satellites + network-satellites for leaf/spine/super-spine switches.
- Various geometries. E.g. 2 planes, 1 for compute and 1 for communication.

## ► Radical new ways of parallel LLM training?

- All-reduce is hard to avoid ...

## AI in The Sky =? Pie in The Sky

Musk: “It’s a *no brainer* for building solar-powered AI data centers in space.”

- Costs savings on construction, (solar) power, and (passive) cooling, and on regulation, NIMBY, lead times, property taxes, ...
- Not included: costs for the astronomy community and society at large (ozon layer!)

Orbital AI inference looks plausible, provided it is confined to individual satellites.

Orbital AI training of frontier LLM models looks daunting,

- Network bandwidths, today’s technologies, falls two orders of magnitude short.

[2026, SpaceX, SEC filing]:

*The timeline for deploying 100 gigawatts of annual compute power to orbit involves unproven or new innovations, and may be difficult or impossible to determine.*

- Cost-effective training in space within 2-3 years looks utterly implausible.  
Even a 10-year horizon looks optimistic to me.

Manuscript to appear on <https://arxiv.org/abs/> soon.