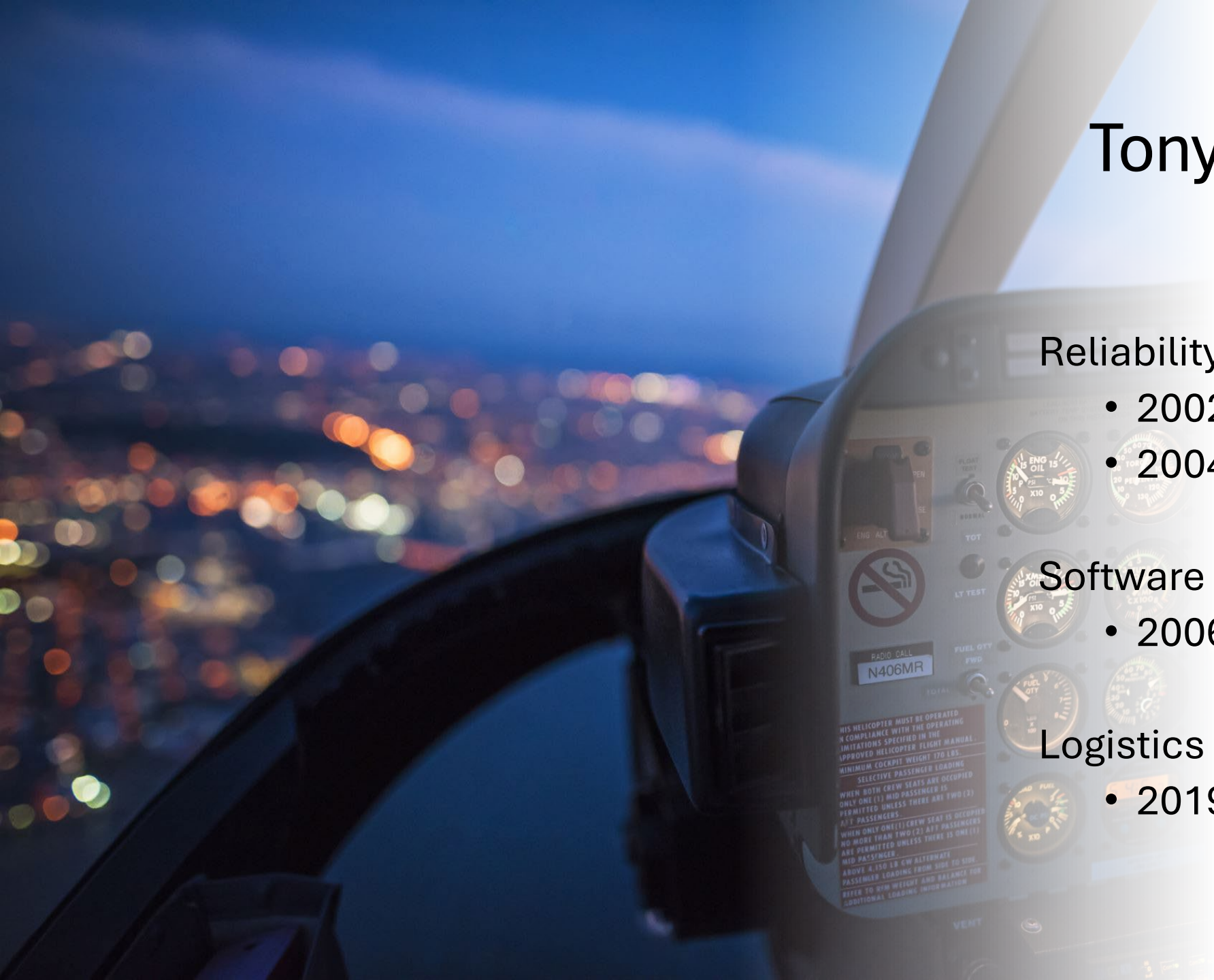
RELIANT TECHNOLOGIES INC.

Reliability and AI

RAM XVI Training Summit



Tony Donatelli

Reliability

- 2002-2003: HIMARS fielding
- 2004-2005: UAV mishaps

Software and Design

- 2006-2019: Army Game Studio

Logistics Engineering

- 2019-2024: LogLab

What is an LLM?

It’s still Machine Learning

Transformers were originally for translation

Pre-training (Baking)

Fine-tuning (Icing)

Commence remixing

And let’s discuss the non-technical

Models I Use

- OpenAI ChatGPT (general use)
- Anthropic Claude (writing)
- Midjourney (images)
- Llama 3 (open source)
- NotebookLM (huge context)
- Grok via X (fun)
- Venice.ai (privacy)
- Replit premium (in-line coding)
- Gamma (presentations)
- Canva (publications)

Best Practices

TELL IT EVERYTHING

STAY OPEN TO ANYTHING

EVERYTHING STARTS AT 70%

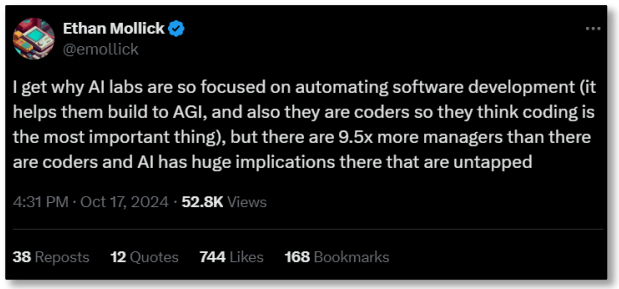
OUTLINE OUTLINE OUTLINE

SITUATIONAL AWARENESS

The Decade Ahead

The Intelligence Age

September 23, 2024



Baking Reliability

Component	Category	Interruption Count	% of Interruptions
Faulty GPU	GPU	148	30.1%
GPU HBM3 Memory	GPU	72	17.2%
Software Bug	Dependency	54	12.9%
Network Switch/Cable	Network	35	8.4%
Host Maintenance	Unplanned Maintenance	32	7.6%
GPU SRAM Memory	GPU	19	4.5%
GPU System Processor	GPU	17	4.1%
NIC	Host	7	1.7%
NCCL Watchdog Timeouts	Unknown	7	1.7%
Silent Data Corruption	GPU	6	1.4%
GPU Thermal Interface + Sensor	GPU	6	
SSD	Host	3	
Power Supply	Host	3	
Server Chassis	Host	2	
IO Expansion Board	Host	2	
Dependency	Dependency	2	
CPU	Host	2	
System Memory	Host	2	

Table 5 Root-cause categorization of unexpected interruptions during a 54-day period of Llama
78% of unexpected interruptions were attributed to confirmed or suspected hardware issue

3.3.4 Reliability and Operational Challenges

The complexity and potential failure scenarios of 16K GPU training surpass those of much larger CPU clusters that we have operated. Moreover, the synchronous nature of training makes it less fault-tolerant—a single GPU failure may require a restart of the entire job. Despite these challenges, for Llama 3, we achieved higher than 90% effective training time while supporting automated cluster maintenance, such as firmware and Linux kernel upgrades (Vigraham and Leonhardi, 2024), which resulted in at least one training interruption daily. The effective training time measures the time spent on useful training over the elapsed time.

During a 54-day snapshot period of pre-training, we experienced a total of 466 job interruptions. Of these, 47 were planned interruptions due to automated maintenance operations such as firmware upgrades or operator-initiated operations like configuration or dataset updates. The remaining 419 were unexpected interruptions, which are classified in Table 5. Approximately 78% of the unexpected interruptions are attributed to confirmed hardware issues, such as GPU or host component failures, or suspected hardware-related issues like silent data corruption and unplanned individual host maintenance events. GPU issues are the largest category, accounting for 58.7% of all unexpected issues. Despite the large number of failures, significant manual intervention was required only three times during this period, with the rest of issues handled by automation.

Sometimes, hardware issues may cause still-functioning but slow stragglers that are hard to detect. Even a single straggler can slow down thousands of other GPUs, often appearing as functioning but slow communications. We developed tools to prioritize potentially problematic communications from selected process groups. By investigating just a few top suspects, we were usually able to effectively identify the stragglers.

One interesting observation is the impact of environmental factors on training performance at scale. For Llama 3 405B , we noted a diurnal 1-2% throughput variation based on time-of-day. This fluctuation is the result of higher mid-day temperatures impacting GPU dynamic voltage and frequency scaling.

During training, tens of thousands of GPUs may increase or decrease power consumption at the same time, for example, due to all GPUs waiting for checkpointing or collective communications to finish, or the startup or shutdown of the entire training job. When this happens, it can result in instant fluctuations of power consumption across the data center on the order of tens of megawatts, stretching the limits of the power grid. This is an ongoing challenge for us as we scale training for future, even larger Llama models.

Meta

The Llama 3 Herd of Models