USATLAS FACE-TO-FACE · THROUGHPUT COMPUTING WEEK 2026

LLM-Assisted Analysis at the UChicago AF

Agentic tools your analysis can use today

Giordon Stark (UC Santa Cruz / SCIPP) · on behalf of the UChicago AF team · 2026-06-09

Slides AI-assisted by Claude (Anthropic · Opus 4.8, 1M) · title illustration AI-generated.

THE WHOLE TALK IN ONE EXAMPLE

“Where do my TopCPToolkit outpu

Your batch jobs emit big **ntuples** and small **histograms**. Where

A generic LLM

“Transfer the outputs back to submit node”, i.e. `$HOME`. The big ntuples blow the 100 GB `/home` quota almost immediately.

An L

big nt
It kno

THE THESIS

The model didn’t get smarter. We gave it **facility context**. That gap, betwe
scheduler, and data, is the key.

GROUNDING · THE VOCABULARY

The vocabulary, fast

Agent

The **engine** doing the work: a loop that uses an LLM (the “brain”) to decide each step, call a tool, read the result, and repeat.

MCP

Model
(Rucio
any ag

Skill

Static context the agent loads on demand: instructions, a reusable recipe (“how we do a pyhf fit here”).

Ager

The ap
writes
Codex

Bundle skills + MCPs + agents + hooks → a **plugin**. Many plugins in one plac
other marketplaces).

RELATED WORK · THE FRAMING

The ecosystem framing (Watts)

> *"It's not about a smarter LLM — it's about smarter infrastructure around it."*
>
> G. Watts, "Beyond Code Generation," CHEP 2026

His framing: data & tools → MCP → skills → agents; grounded in primary sources; "we need a **pip-install**."

This talk: one facility's instance, running now, and the claim that **facility context** is the decisive layer, built to port.

Watts, CHEP 2026: indico.cern.ch/event/1471803/contributions/6969164

01

THE PRESENT

What an analyzer can use today

WHAT YOU CAN USE TODAY · MANAGED SETTINGS

You bring the agent; we ship the ru

You install your own harness on the login nodes; we don’t force one

Our config management (Puppet) ships system-wide **managed settings**: a curated **allow-list** of safe HEP commands + the **ATLAS env**

Any harness that reads them inherits the facility’s guardrails

manage

```
{
  "
  ]
  "
  }
}
```

WHY IT MATTERS

The facility, not the user, decides what an agent may do here.

OPEN QUESTION

Ship this with the facility (Puppet) or via a centralized marketplace? We’ll come back to it.

Shipped system-wide by our config management (Puppet) to `login01-04` · live since 2026-05

WHAT YOU CAN USE TODAY · THE KEY IDEA

The facility teaches the agent abou

Auto-loaded into every session: `/etc/claude-code/CLAUDE.md`

Path	Quota	Use for
`/home/$USER`	100 GB, backed up	Code, scripts, condor files
`/data/$USER`	5 TB, not backed up	ROOT files, datasets
`/scratch`	node-local	Ephemeral, copy out before exit

```
# live monitor (DO NOT use: watch condor_q)
condor_watch_q
# why is a job held?
condor_q -hold
# XCache-optimized reads (SITE_NAME=AF_200)
rucio list-file-replicas <scope>:<name> --protocol root
```

TH
A
ti
w

S

S

D

p

T

WHAT YOU CAN USE TODAY · DISTRIBUTION

The USATLAS marketplace: installi

```
> /plugins marketplace add usatlas/marketplace
```

Three plugins (each bundles skills, subagents & hooks) at github.com/usatlas/ma

analysis-facilities

Facility skills: HTCondor, JupyterLab, XCache, Rucio, ServiceX, Coffea-Casa, Triton (for UChicago, BNL, SLAC).

atlas

5 subagents + 25+ skills: Rucio Open-Data MCPs, pyhf, cabine TRExFitter, TopCPToolkit, Fastl Scikit-HEP.

PLUGINS INSTALL IN ONE LINE; THE TOOLS DON'T

The agent know-how adds instantly; the tools come via **setupATLAS** + **pixi**. No `/cvmfs` ? **pixi + conda-forge** covers most — incl. **ARM builds** & **supply-chain security** (e.g. pixi **dependency cooldowns**) via HEP Packaging Coordination (cf. Feickert, CHEP 2026).

WHAT YOU CAN USE TODAY · DATA & METADATA MCPS

Find datasets & metadata by asking

`rucio-mcp` : dataset access (find, inspect, replicas, download)

`ami-mcp` : metadata & cross-sections

also `atlasopenmagic` , glance via `stare`

Run it locally (your x509)

```
pixi exec rucio-mcp serve --read-only
pixi exec ami-mcp serve
uvx atlasopenmagic-mcp serve
```

`rucio-mcp` alone exposes **50+ tools**; `--read-only` blocks every write.

Or t

```
# mu
rucio
  · a
```

Browser
yet (FTS,

STILL BEING FIGURED OUT

OIDC UX still varies by VO: token lifetime & refresh differ; audiences differ (atlas/dune: no custom issuer; cms/escape: extra token-exchange). And on a hung call, who times out: MCP LLM?

Auth reuses grid creds (x509) or Keycloak OIDC · OAuth bridge docs · running workflows/jobs is

WHAT YOU CAN USE TODAY · JUPYTER MCP

Drive your AF notebook from anyw

An MCP server runs **inside your JupyterLab pod**. Any agent that reaches it (Claude Code, Claude Desktop, even your **phone**) can `insert_execute_code_cell`, `read_cell` (incl. plots), `execute_code` in the live kernel, `use_notebook`.

≠ **Jupyter AI** (agent inside the notebook server), which may lose out to the VS Code workflow many users prefer.

```
claude mcp add jupyter --transport http \
  .../user/<you>/mcp --header "Authorization: Bearer
<tok>"
```

🔑 **THE CONVENIENCE/SECURITY TRADEOFF**

Servers have finite lifetimes (equitable sharing) → you re-mint URL + token each respawn. Idea: a Keycloak-OIDC reverse-proxy → one token, pick your server in-browser (but one at a time).

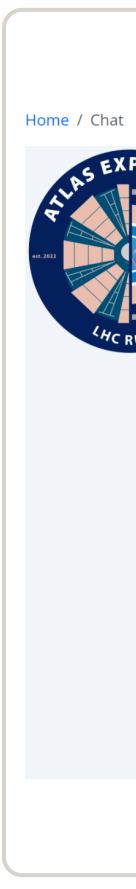
Jupyter MCP streamable-HTTP extension · full-size demo video: Google Drive.

WHAT YOU CAN USE TODAY · OPENWEBUI

A web chat that already knows the

Not everyone wants a terminal. **OpenWebUI** is a browser chat at af.uchicago.edu/chat, backed by a facility knowledge base.

- Zero install, just a URL
- Answers based on AF docs & capabilities
- Same knowledge, friendlier UI/UX



REAL EXCHANGE · USER `XJU` , FEB 2026

Q: "Are there NVIDIA Triton models available at AF?"

A: Yes. Triton in the AF k8s cluster, serving from CVMFS + an S3 repo (`s3://triton-models/<user>/`); upload yours, then ask admins to enable it.

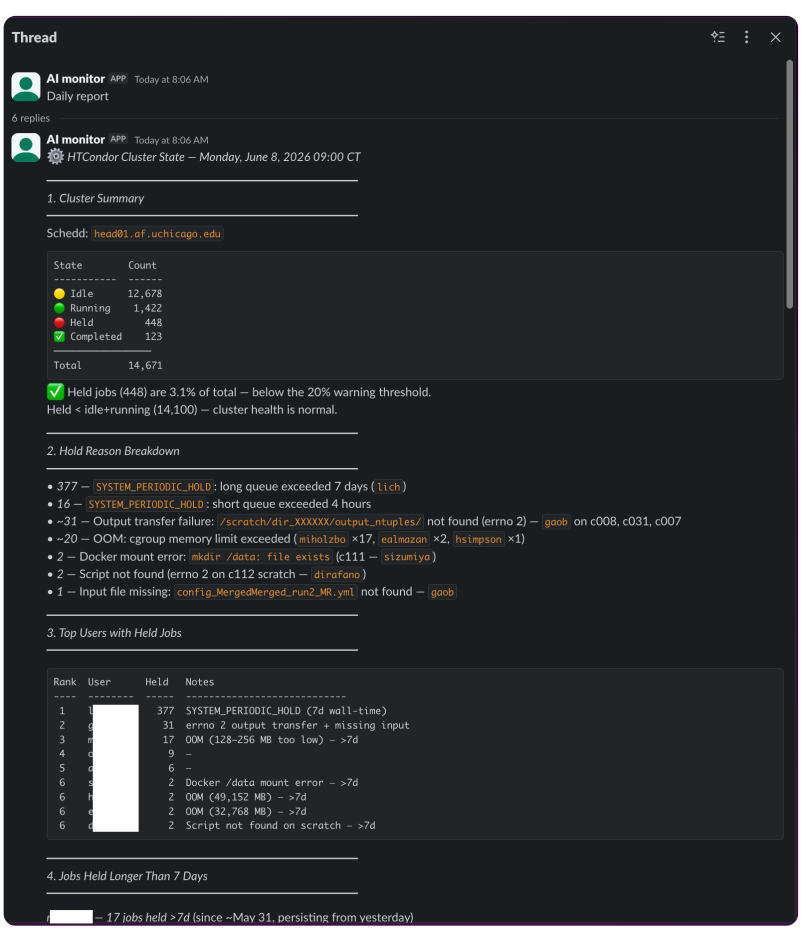
"this... is not awful and it is correct" — Giordon

Even

Assistant built by Ilija Vukotic (slides) · it also pointed `xju` to xAOD/coffea examples · Triton→a

WHAT YOU CAN USE TODAY · AGENTS ON YOUR BEHALF

An agent watches the cluster for yo



Thread

AI monitor APP Today at 8:06 AM
Daily report

6 replies

AI monitor APP Today at 8:06 AM
HTCondor Cluster State — Monday, June 8, 2026 09:00 CT

1. Cluster Summary

Schedd: `head01.af.uchicago.edu`

State	Count
Idle	12,678
Running	1,422
Held	448
Completed	123
Total	14,671

✅ Held jobs (448) are 3.1% of total — below the 20% warning threshold.
Held < idle+running (14,100) — cluster health is normal.

2. Hold Reason Breakdown

- *377* — `SYSTEM_PERIODIC_HOLD`: long queue exceeded 7 days (`lich`)
- *16* — `SYSTEM_PERIODIC_HOLD`: short queue exceeded 4 hours
- *~31* — Output transfer failure: `/scratch/dir_XXXXXX/output_ntuples/` not found (errno 2) — `gaob` on c008, c031, c007
- *~20* — OOM: cgroup memory limit exceeded (`miholzbo` ×17, `ealmazan` ×2, `hsimpson` ×1)
- *2* — Docker mount error: `mkdir /data: file exists` (c111 — `sizumiya`)
- *2* — Script not found (errno 2 on c112 scratch — `dirafano`)
- *1* — Input file missing: `config_MergedMerged_run2_MR.yml` not found — `gaob`

3. Top Users with Held Jobs

Rank	User	Held	Notes
1	l[redacted]	377	SYSTEM_PERIODIC_HOLD (7d wall-time)
2	g[redacted]	31	errno 2 output transfer + missing input
3	m[redacted]	17	OOM (128–256 MB too low) — >7d
4	c[redacted]	9	—
5	d[redacted]	6	—
6	s[redacted]	2	Docker /data mount error — >7d
6	H[redacted]	2	OOM (49,152 MB) — >7d
6	e[redacted]	2	OOM (32,768 MB) — >7d
6	d[redacted]	2	Script not found on scratch — >7d

4. Jobs Held Longer Than 7 Days

[redacted] — *17 jobs held >7d* (since ~May 31, persisting from yesterday)

Every mo
report to
users, job

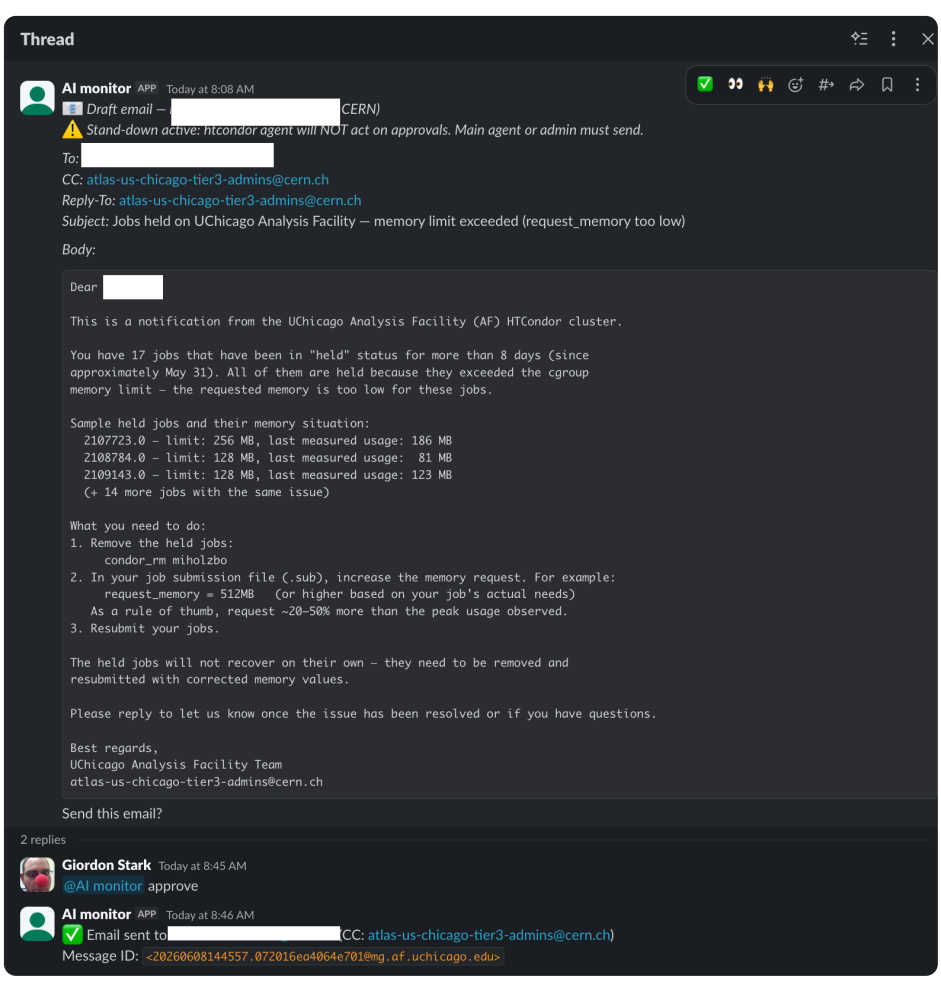
> Today:
> Holds a

No HTC

Two lanes on **OpenClaw**: **Shannon** = privileged, trust-earned runbooks (this bot, human-gated)
HTCondor MCP exists (Bockelman); deeper → next talk.

WHAT YOU CAN USE TODAY · AGENTS ON YOUR BEHALF

...then drafts your fix, you approve



When you
specific f
vs the lim
`data` .

> **HUM**
> Drafted
> **autono**

⚠ Good ad
`c113` , `/hom`

> **THIN**
> How do v
> context?
> agent ca

02

THE ARCHITECTURE

Why it works

THE CORE INSIGHT

Swap the model freely: context ma

Model
Opus · GPT-5 · **local / NRP-hosted open weights**
→ cost control

Harness
Claude Code · Codex · Copilot

An assistant that's actually
routes ntuples to `/data`, not

The first two blocks are **interchangeable**: pick any model (or self-hos
from the **third**: the managed `CLAUDE.md`, the marketplace plugins, the

Pragmatically too: model & harness evolve fast, so we can't run a cluste

Self-hosting option for open-weight models: NRP-hosted LLMs (nrp.ai/llms).

THE CORE INSIGHT · THE FLIP SIDE

No context? It should stay quiet.

⚠ WHAT ACTUALLY HAPPENED

Our cluster agents confidently recommended **Lustre/MDS tuning** for a file
context. Plausible, fluent, and wrong.

"If we're supposed to rely on the agents, they need to be accurate — otherwise they can't be trusted."

Judith, `#analysis-facility`

"If th reco sugg

Farna

The rule we landed on: **if it isn't grounded in real facility context, it stays** makes the agent useful — it's what makes it safe to trust.

THE ARCHITECTURE · ABSTRACTION

The agent speaks intent, not imple

What the physicist & ager

(meta)data_tool
“get me this dataset + its cross-section”

transform_tool
“turn DAODs into histograms”

“ru

What the **facility** wires underneat

rucio-mcp · ami-mcp · Open Data

coffea · uproot · FastFrames · ServiceX · Athena

HTCon
· REAN

The agent never names “coffea” or “condor.” It asks for an outcome; the fa
makes the same agent portable across facilities.

THE ARCHITECTURE · THE “ELWOOD” VOCABULARY

Reasoning engine & playbook

Reasoning engine

The framework: orchestration, tool routing, execution, guardrails. Model- and experiment-agnostic.

shared written once

P

Every
the sy
examp

swap

Think of it as a sports team

A shared **glossary** keeps the team’s language consistent. (“Elwood” is our internal name, nothing public yet.)

Ima

THE ARCHITECTURE · PORTABILITY, CONCRETELY

Same engine, different playbook

Only the playbook changes per facility/experiment. The reasoning engi
is usually `data_tool` :

playbook/atlas @ UChicago AF

```
data_tool:   rucio-mcp     # grid
transform:   coffea, FastFrames
batch_tool:  htcondor
inference:   pyhf, TRExFitter
```

playb

```
data_
trans
batch
infer
```

> The per-tool **skills are facility-independent** (written once, reused everywh
> playbook, not a whole stack. **The goal: stop re-developing the same ager**

You've already met the ATLAS playbook in pieces: the managed `CLAUDE.md` , the marketpla

Still open: how to link & reuse others' skills without copying them in: usatlas/marketplace#60

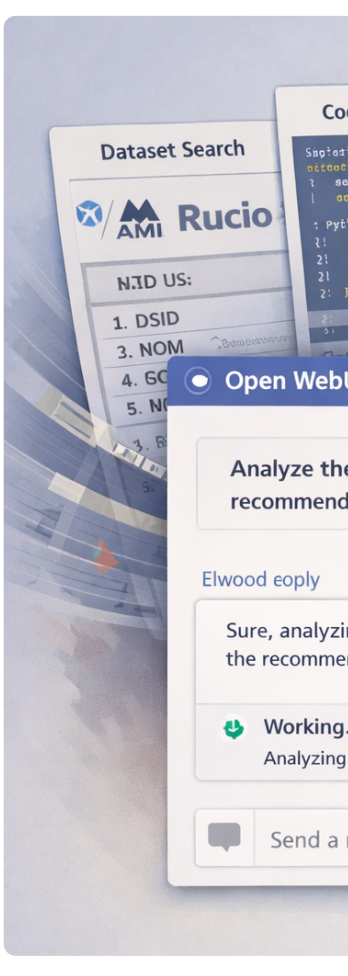
THE ARCHITECTURE · WHERE IT'S HEADING

One conversation, the whole analy

The target: the analyzer describes the physics; the agent runs the loop: find data, generate code, submit jobs, recover from errors, make histograms, and fit, re-running as needed.

Every box here is a tool we've shown today. Stitching them into one supervised loop is the work ahead.

"Elwood" is our internal project name, nothing public yet.



03

THE PORTABLE FUTURE

Toward HL-LHC analysis facilities

THE PORTABLE FUTURE · MCP GRANULARITY

One MCP, or many? Three topologi

Live today: `rucio-mcp.af.uchicago.edu/site/{atlas·escape·cms·`

what we run

One server, many sites

you → rucio-mcp /site/atlas·escape…

✓ one deploy; add `rucio-atlas` , `rucio-escape` separately in your harness

⚠ per-VO auth inside one process

alt

One server per site

rucio-atlas rucio-escape

✓ clean isolation; per-site auth & sc

⚠ N servers to run & maintain

Same agent, same tools across sites; only the per-site auth/playbook differs. Granularity vs iden

LET'S DISCUSS · THE OPEN QUESTIONS

What should we decide together?

Granularity & identity

One MCP per service, or a single gateway delegating identity? (FastMCP allows one auth provider/server; PandaMCP delegates to PanDA.)

Where do knowledge & live?

Facility knowledge: shared **market the facility** (Puppet). And tool skills `TopCPToolkit` skill in the marketp **the framework**? Today: scattered.

What may an agent do, and how sandboxed?

Write/submit/email need a human gate. We isolate with **k8s** pods; no-k8s sites → bubblewrap?

Who pays for inference

At HL-LHC scale: hosted frontier m hosted open weights on facility GP

My bet: the model and harness are the easy part. **Context, id**

WHAT'S BUBBLEWRAP?

The sandboxing tech behind Flatpak: it isolates a process from the rest of the Linux syste explicitly grant it. A lightweight way to fence in an **agent's tool execution** on facilities wit

THANK YOU

Generic LLM → facility-aware colla

Swap the model and the harness freely. **The facility context is what makes it useful here**, and it ports to the next AF.

A team effort at the UChicago AF

Ilija Vukotic — OpenClaw/Shannon, ES MCP

Rob Gardner — vision, marketplace, Genesis

Aidan Rosberg — RP1 (Infra-as-Config), core dev/maintainer

Farnaz Golnaraghi — hardware ops: storage

Fengping Hu — Kubernetes, Keycloak, Jupyter AI

Judith Stephen — HTCondor expertise, runbooks

David Jordan — hardware ops: networking & Kubernetes

→ **What's next:** port these lessons (incl. agentic AI) to the Open Data Facility (ODF) and RP1.