AI DRUG DISCOVERY & DEVELOPMENT SUMMIT, BOSTON, 29 OCTOBER 2026

Requested vs. Used: Sizing Workloads Before They Run

Ismaeel Bashir, Co-founder & CEO, Expanse

expanse

Compute certainty.



“

It’s an art form. When a job crashes, double the resources and try again until it works.

How I was taught to size jobs on a supercomputer

THE SUBMIT SCRIPT

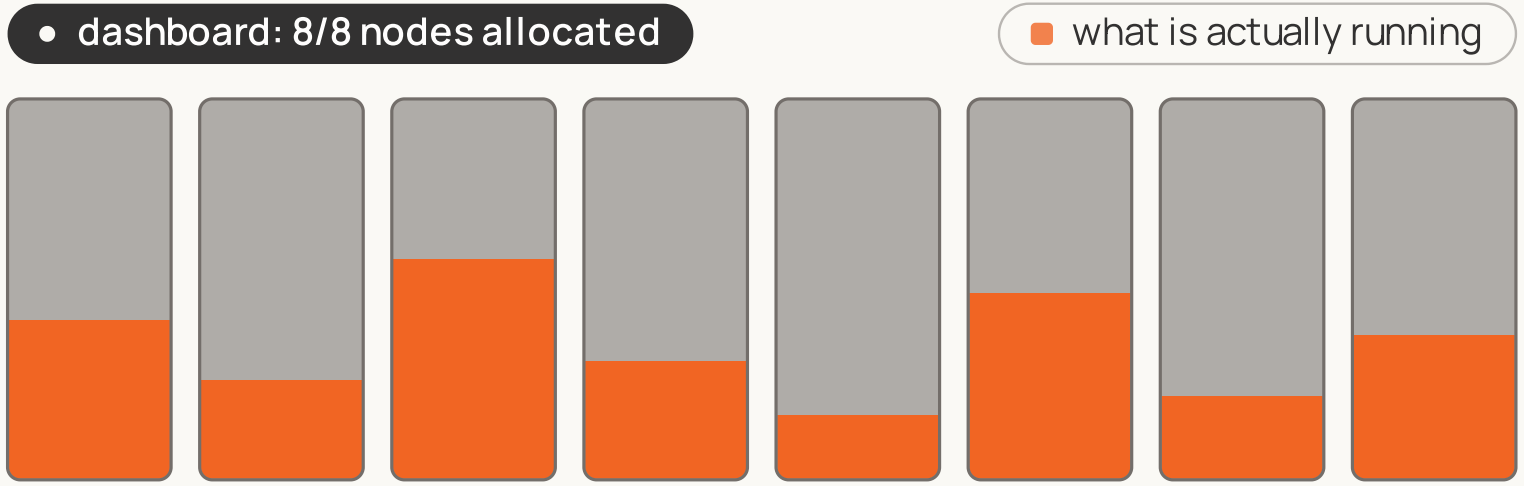
Every job on every cluster starts with a guess

```
#!/bin/bash
#SBATCH --job-name=md_production
#SBATCH --mem=128G
#SBATCH --time=24:00:00
#SBATCH --nodes=4

srun gmx_mpi mdrun -deffnm prod
```

WHAT THE DASHBOARD COUNTS

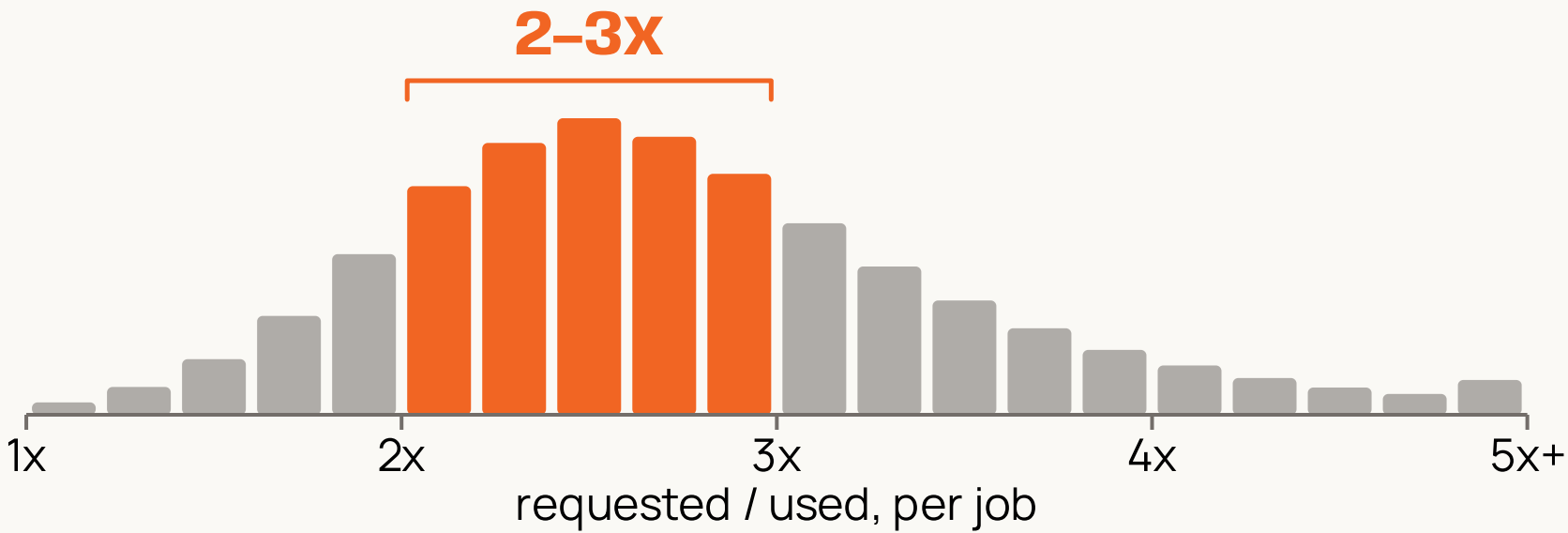
A cluster can be fully allocated and half idle



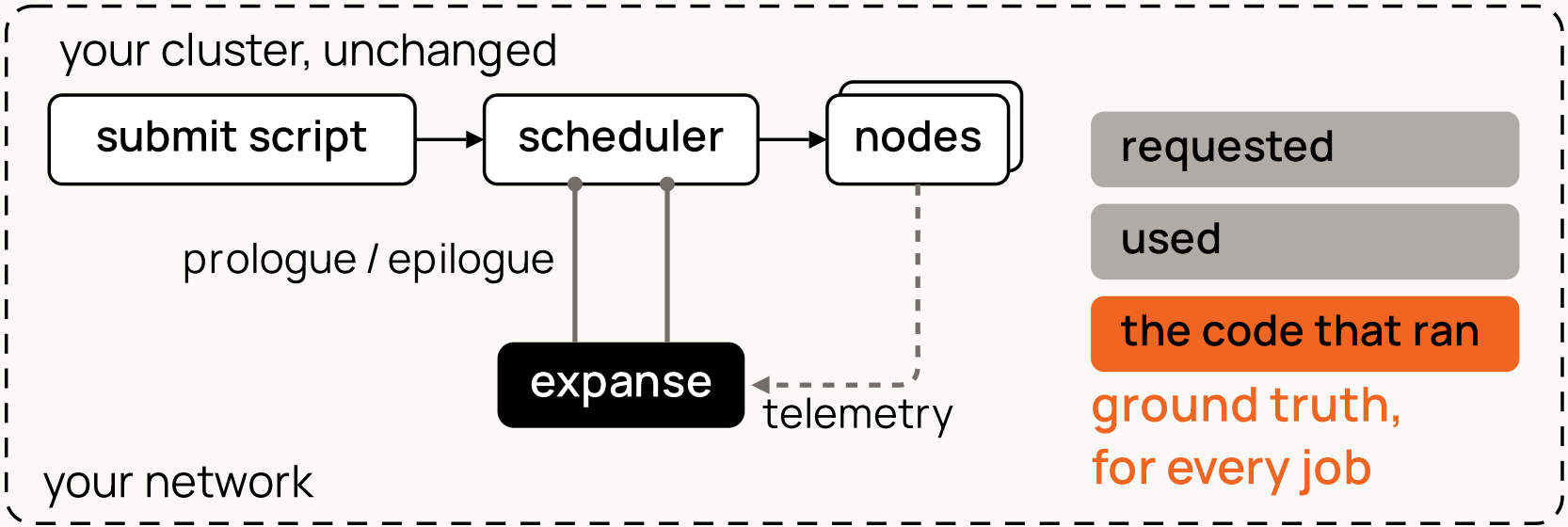
the gap is inside the allocation, not between the nodes

A MONTH OF JOBS, TWO PRODUCTION CLUSTERS

Two to three times more reserved than used



We measured for a month without changing anything

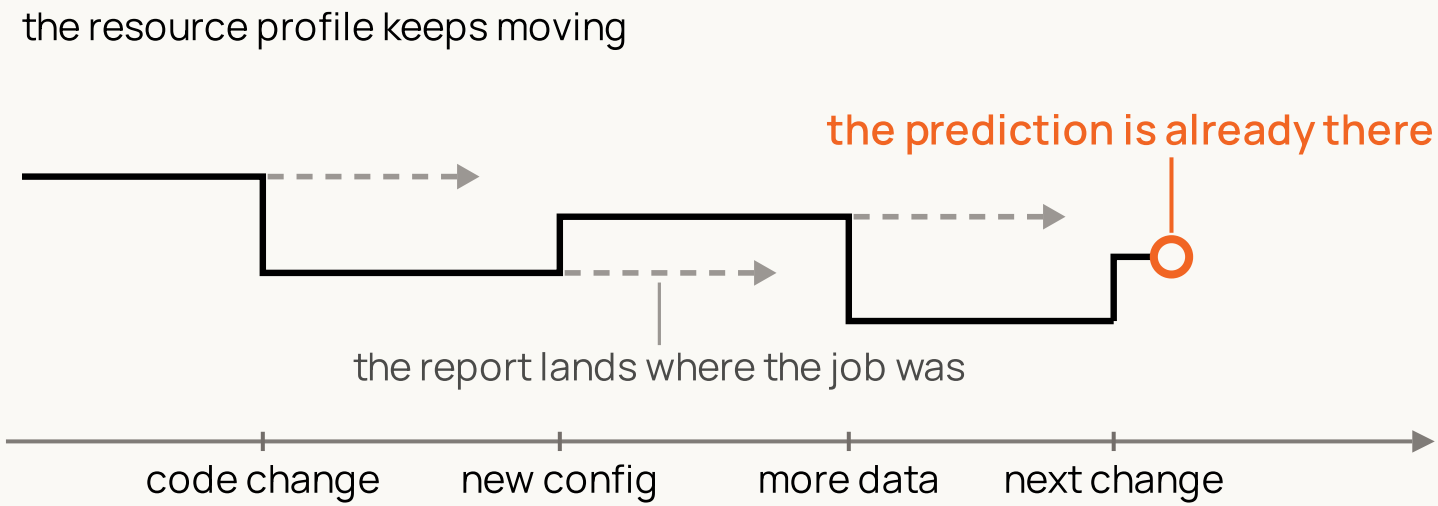


ONE MONTH, ONE CLUSTER

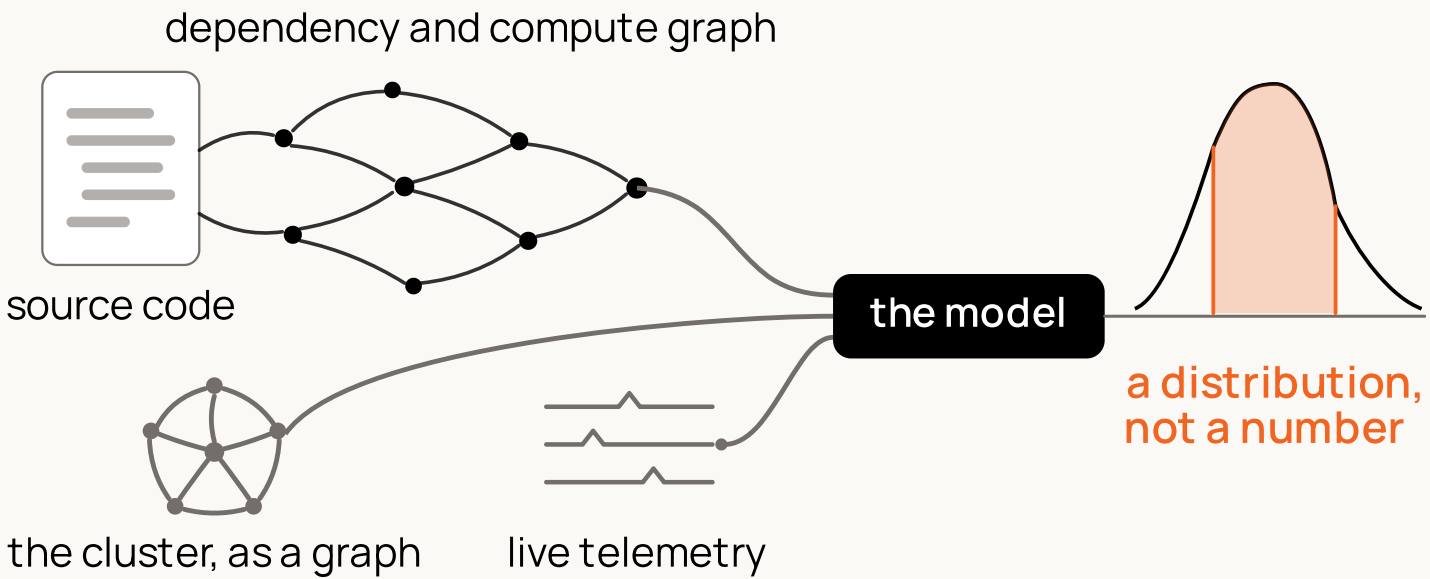
$8M

of compute capacity identified, with nothing bought and nothing changed

A report chases the job. A prediction meets it.



The model reads code as a graph, not as text

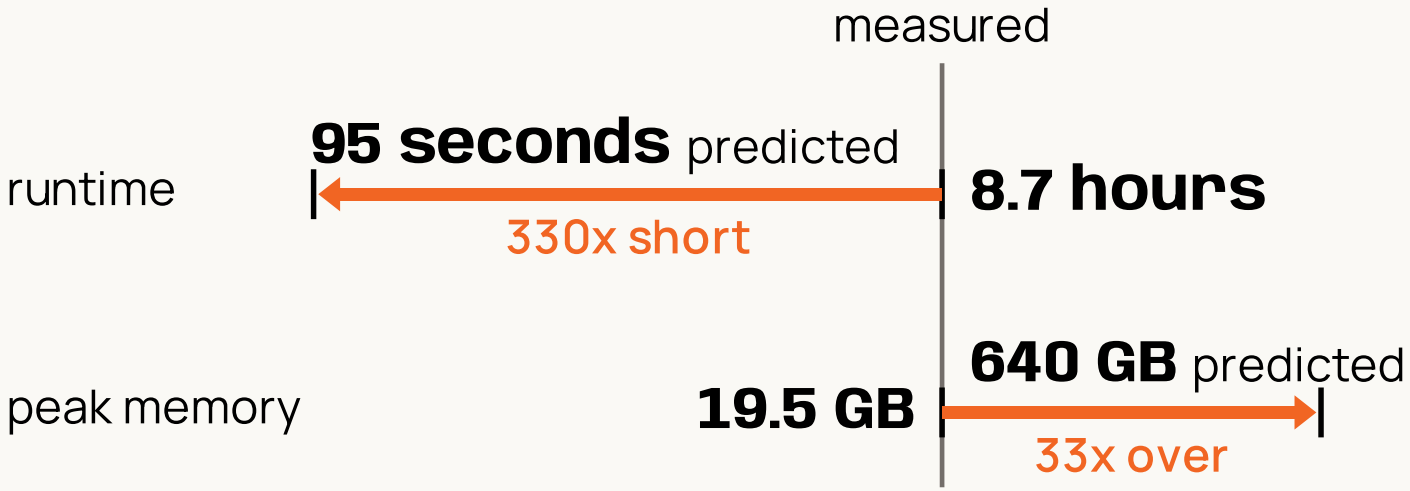


BEFORE THE JOB RUNS

The prediction arrives before the job is submitted

```
$ expanse analyse md_production.sbatch
runtime     31 h              [28 - 34]
memory      9 GB              [7 - 11]
cores       128               [112 - 144]
failure     0.87              time limit, 24 h requested
fix         --time 36:00:00 what similar jobs needed
```

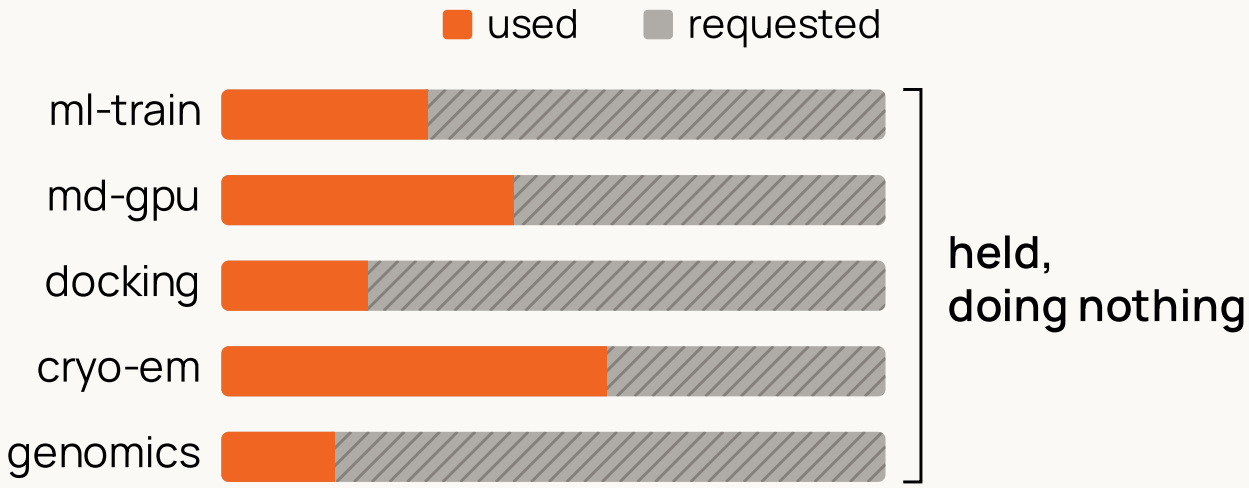
Frontier models are wrong in both directions



Expanse benchmark, 2026. Eight frontier models, 102 HPC programs, full source given.

FOR WHOEVER OWNS THE CLUSTERS

The operator finally sees where the capacity goes



Same submit script. Same scheduler. More output.

$8M IN ONE MONTH, ON ONE CLUSTER

What’s the number on yours?

Compute certainty.

Scan to book a call

expanse

expanse.sh ismaeel@expanse.sh