AI and Data Center Resource Usage

New Mexico Science, Technology, and Telecommunications Committee

Monday, September 30, 2024

Patrick G. Bridges

Director, Center for Advanced Research Computing

Professor, Department of Computer Science

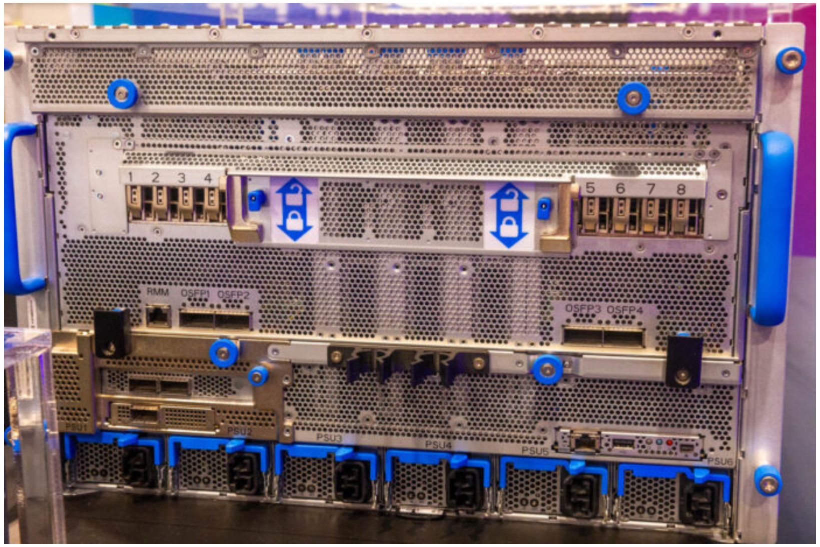
University of New Mexico

Data Centers Everywhere (including NM)

- **New Mexico is an attractive place to build data centers**
  - Abundant solar and wind power, potentially good cooling options
  - Inexpensive property costs, few pesky natural disasters
- **Been examining data center options for NM Universities**
  - Advanced computing research and education at UNM
  - AI systems for NM Artificial Intelligence Consortium
  - Need research and training systems/facilities for New Mexico students
- **Lots of complex tradeoffs in powering and cooling data centers, particularly for AI workloads**
- **AI computing systems are very *dense,* take lots of power to make run, and generate a lot of heat in a very small space**

What kinds of resources to AI clusters use?

- **What goes into an AI data center?**
  - Racks of computers with GPUs/TPUs are the AI workhorse
  - Storage for holding data sets, network gear, standard computers
  - Data
  - Power
- **What goes out of an AI data center?**
  - Data
  - Heat
- **AI compute systems are the big resource consumer here**
  - Thousands of compute servers + accelerators (GPUs, etc.)
  - Example: Microsoft Eagle (2023): 1800 nodes, 14,400 total GPUs
  - Microsoft says they are deploying the equivalent of 5 of these per *month*



https://www.servethehome.com/microsoft-azure-eagle-is-a-paradigm-shifting-cloud-supercomputer-nvidia-intel/

Two Main AI Datacenter Resource Drivers

- **IT Power: The power that the equipment itself draws to run**
  - Each the 1800 NBv5 systems in Microsoft Eagle uses ~5 kilowatts power, but some emerging technologies (dedicated AI accelerators) may help
  - For reference: the average NM home your house uses ~700 watts
  - Power Usage Efficiency (PUE): $\frac{Facility\ Power}{IT\ power}$ (1.0 ideal, 1.2 is good today)
  - Microsoft Eagle: 10 MW of IT power, assume 12 MW of facility power
- **Cooling: Getting the generated heat out of the facility**
  - ***This the extra resource utilization to pay special attention to!***
  - The inexpensive, efficient ways to do this typically involve water
  - Refrigerated air uses minimal water but leads to a PUE > 2.0!
  - Water Usage Efficiency (WUE): $\frac{Water\ Consumption}{kWh\ of\ facility\ power}$ (< 0.5L / kW is good)
  - 12MW at 0.5 WUE is about >13,000,000 gallons of water per year
  - Result: $250,000 water bill, but saves $Millions in power

Lots of options for cooling

- **Cooling systems are built on two main heat transfer loops**
- **Cooling loop – get heat out the room/equipment**
  - Air-cooling – blow cool air (65°F) to computer components
  - Direct liquid to chip – pump enclosed water (75°F) past hot components
  - Immersion - put **entire system** in very warm (90-120°F degree) special non-conductive coolant, requires specialty hardware and handling
  - Modest resource usage here – fans and pumps
- **Heat rejection loop - Get the heat out of the coolant loop**
  - *This is where resources are used for cooling!*
  - Dry cooling requires coolant loop target temp 10° warmer than ambient
  - Can supplement with evaporative/AC chillers when too hot outside
  - Chilled water loops can be directly evaporatively cooled
- **Lots of additional innovations in the space**
  - Water economizers, mixed refrigerant/evaporative systems (Sandia), etc.
  - Ways to recycle water or even the system heat

Takeaways

- Complex technology, conservation, and economic tradeoffs with modern AI data centers
- Current economics favor using water for large-scale cooling of data centers
- Options are emerging (DL2C, Immersion, Improved cooling technologies) to improve cooling resource efficiecy
- UNM working with experts on best options for our own data center AI research, education, and workforce needs