

Virginia Tech – Facilities, Equipment and Other Resources

Virginia Tech's Advanced Research Computing (ARC) is a unit within the Division of Information Technology and provides centralized research computing resources which are available to any Virginia Tech researcher. ARC provides cutting-edge high-performance computing and operates 5 clusters with 632 compute nodes, 64,736 CPU cores, 360 TB RAM, 514 GPUs, and over 11 PB of storage. Currently available high performance computing (HPC) systems include:

1. **TinkerCliffs:** flagship general-purpose CPU and GPU cluster. This cluster has 288 nodes with 36,864 AMD Rome CPU cores, 16 nodes with 1,536 Intel Xeon CPUs, 8 high-memory nodes with 1 TB RAM each, 14 nodes with 8 NVIDIA A100-80GB GPUs each, and 6 nodes with 8 NVIDIA H200-141GB GPUs each. Nodes are connected via HDR InfiniBand offering 100 Gbps throughput.
2. **Owl:** high-speed, water-cooled CPU and GPU cluster. This cluster has 160 nodes each with 96 AMD Genoa CPU cores and 768 GB RAM, 4 nodes with 64 AMD Milan CPU cores and 512 GB RAM, 12 high-memory nodes each with either 3, 4 or 8 TB RAM, and 4 nodes with 8 NVIDIA B200-180G GPUs each, interconnected with HDR InfiniBand running at 200 Gbps.
3. **Falcon:** GPU-based cluster made up of 106 compute nodes with a total of 128 NVIDIA A30, 80 NVIDIA L40S, 80 NVIDIA V100, and 18 NVIDIA T4 GPUs. Nodes are connected via NDR InfiniBand offering 200 Gbps throughput or 10 Gb Ethernet (V100 and T4 nodes).
4. **CUI:** This is a dedicated cluster for ITAR or Export Controlled software/data. This cluster has 10 CPU nodes each with 64 CPUs and 512 GB RAM, and 2 GPU nodes each with 64 CPUs, 2 TB RAM, and 8 NVIDIA A100-80GB GPUs. Access is restricted to US persons.
5. **Biomed:** This is a dedicated cluster for biomedical research needing NIST 800-171 compliance. This cluster has 2 CPU nodes each with 64 CPUs and 512 GB RAM, and one GPU node with 64 CPUs, 2 TB RAM, and 8 NVIDIA A100-80GB GPUs.

Centralized storage provides over 11 PB of long-term bulk "project" storage, as well as "scratch" storage for workflows with demanding I/O needs.

ARC employs 16 FTEs and 3 GRAs to support the infrastructure and provide user support. ARC provides services to accelerate discovery and enhance the use of these systems. ARC has a staff of computational scientists who are available to consult with researchers to provide expert advice on the selection of systems for their workloads, optimizing workflows and code bases, and actively engaging in collaborative research. For extended engagements, they are also able to participate as named personnel on sponsored projects. Daily office hours are hosted by a team of graduate assistants who also provide most of the support for system usage via ARC Helpdesk tickets.

ARC provides self-hosted LLMs for research, education, and administrative use, with several ways to access them. Users can interact through a web interface at <https://llm.arc.vt.edu> or integrate programmatically via the API at <https://llm-api.arc.vt.edu>. Dedicated LLM environments are available through Open OnDemand at <https://ood.arc.vt.edu>. ARC supports custom deployments using vLLM and llama.cpp with Slurm, as well as integration with IDEs using APIs.

ARC's Visionarium Lab provides an array of visualization resources, including the VisCube, an immersive 14.7' x 14.7' x 9.2' three-dimensional visualization environment. The VT Visionarium provides nearly 86 million pixels, 4 billion triangles-per-second and 22 TB/s of GPU bandwidth.

ARC resources leverage Virginia Tech's network connectivity, and network. Virginia offers access to advanced national networks, including ESnet, Internet2, and Mid Atlantic Crossroads.