State of OSG

Frank Würthwein
OSG Executive Director
UCSD/SDSC

July 8th 2024

Where do we want to go next with OSG?

~~State of OSG~~

Frank Würthwein
OSG Executive Director
UCSD/SDSC

July 8th 2024

SDSC SAN DIEGO SUPERCOMPUTER CENTER

There is a lot of change and opportunity

What if any of these do we want to go after together?

Some Examples

- ”Democratizing Access”
- Integrated Research Infrastructure
- NAIRR Pilot
- FASST

"Democratizing Access"

(see my talk on Tuesday morning)

… or … where do we want growth to come from?


Arctic Ocean
Svalbard
Greenland
Iceland
Sweden
Norway
Russia
Canada
North Atlantic Ocean
North Pacific Ocean
Spain
Italy
Kazakhstan
ASIA
Turkmenistan
China
Japan
Tunisia
Iraq
Iran
Algeria
Libya
Egypt
Oman
Bangladesh
Cuba
Mali
Niger
Senegal
Sudan
Nigeria
AFRICA
Somalia
Maldives
Indonesia
Papua New Guinea
Venezuela
Ecuador
Tanzania
Angola
Namibia
Madagascar
South Africa
Chile
Uruguay
South Pacific Ocean
South Atlantic Ocean
Indian Ocean
AUSTRALIA
OSG Site
245 Sites, 151 Institutions

My Opinion

… or … where do I want growth to come from …

Problem Statement

- OSG requires the institution to operate a batch cluster and/or a storage cluster in order to join the OSG
- Many (?) of the ~4,000 accredited institutions of higher learning may not have the staff to do so.
  - Can't afford them
  - Can't recruit them given location
  - Doesn't make sense to sustain them given size of institution

Can we reduce the total cost of ownership of compute & data infrastructure ?

Integrated Research Infrastructure

- The DOE is putting some focus on “Integrated Research Infrastructure”
- We will hear about this more later today as part of HPDF presentation

Is there something we want to push for together ?

Do we have something to contribute?

NAIRR Pilot

- At this point there is a monthly allocations call to hand out resources.
- Are you taking advantage of this?
- Should we do anything together here?

Frontier AI for Science, Security, and Technology

DOE AI Program lobbied for by Rick Stevens (ANL)

ASCAC Presentation by Rick Stevens

FASST Targets

- **Data effort must produce tokens on schedule** or whole effort will be rate limited on data preparation
  - Labor and inference intensive ⟹ **100 T tokens in first few years**
  - Common data APIs are needed, but not waste time on unneeded sw/standards
- **To train ~10 Frontier FMs per year will require building out of significant AI training resources to avoid cannibalizing LCFs**
  - Need 10x current Exascale AI flops in next few years ⟹ **200 AI EFs (a few sites)**
- **Inference hardware capacity is critical**
  - Need to serve models/apps for development and production
  - Will need thousands of inference servers ⟹ **200 AI EFs (deployed at ~10 sites)**
- **Large increase in staff are needed across the FASST program**
  - Much of the work in building and deploying FMs and applications is "engineering" and a project framework is needed to both manage to schedule and to integrate the hundreds of activities ⟹ **2000 FTEs**
- **Applications need to be deployed to get productivity boost**
  - Applications development should start now with open models and swap FMs as better ones become available ⟹ **100 frontier AI based applications**
  - Modular architecture with plug-in APIs are needed to avoid silos

Five-year Sketch of Data Preparation for FMs

- Accelerators: 100 Trillion Tokens
- Biology: 35 Trillion Tokens
- Chemistry: 35 Trillion Tokens
- Climate: 90 Trillion Tokens
- Computer Science: 3 Trillion Tokens
- Cosmology: 100 Trillion Tokens
- Energy Systems: 63 Trillion Tokens
- Fusion Energy: 100 Trillion Tokens
- HPC codes: 12 Trillion Tokens
- Manufacturing: 100 Trillion Tokens
- Materials: 60 Trillion Tokens
- Mathematics: 42 Trillion Tokens
- Nuclear Physics: 80 Trillion Tokens
- Particle Physics: 80 Trillion Tokens
- Reactors: 100 Trillion Tokens

1000 Trillion Tokens over 5 years?

GPT4 trained on ~15T tokens
Llama3 trained on ~15T tokens

FY26: $100M, 10 Trillion Tokens - Begin with organizing and curating datasets in text or narrative form for AI model training. Initial focus areas include:
- Mathematics: 2 Trillion Tokens
- Computer Science: 3 Trillion Tokens
- HPC codes: 2 Trillion Tokens
- Energy Systems: 3 Trillion Tokens

FY27: $150M, 50 Trillion Tokens - Expand data curation efforts to enhance AI model training capabilities. Add datasets for:
- Biology: 10 Trillion Tokens
- Chemistry: 10 Trillion Tokens
- Materials: 10 Trillion Tokens
- Energy Systems: 10 Trillion Tokens
- HPC codes: 10 Trillion Tokens

FY28: $250M, 150 Trillion Tokens - Further enhance curated datasets to support a broader range of AI applications, preparing for complex AI challenges. Include data for:
- Particle Physics: 30 Trillion Tokens
- Nuclear Physics: 30 Trillion Tokens
- Climate: 40 Trillion Tokens
- Biology: 25 Trillion Tokens
- Chemistry: 25 Trillion Tokens

FY29: $300M, 300 Trillion Tokens - Sustain and expand dataset curation and maintenance to support continuous AI model development. Integrate datasets for:
- Fusion Energy: 50 Trillion Tokens
- Accelerators: 50 Trillion Tokens
- Materials: 50 Trillion Tokens
- Particle Physics: 50 Trillion Tokens
- Nuclear Physics: 50 Trillion Tokens
- Climate: 50 Trillion Tokens

FY30: $400M, 490 Trillion Tokens - Continuously manage and expand curated datasets to enable the development of domain-specific models and synthetic data applications. Finalize with datasets for:
- Cosmology: 100 Trillion Tokens
- Reactors: 100 Trillion Tokens
- Manufacturing: 100 Trillion Tokens
- Fusion Energy: 50 Trillion Tokens
- Accelerators: 50 Trillion Tokens
- Energy Systems: 50 Trillion Tokens
- Mathematics: 40 Trillion Tokens

Key FY26 Deliverables:

- 6 operational AI hubs
- 3 domain-specific foundation models trained on initial curated datasets
- 10 DOE AI FM applications developed and deployed
- Suite of curated datasets to enable further model development
- "20 AI EF" systems deployed
- Upgraded compute infrastructure to support model training
  - Supporting 1,000 DOE active scientific/engineering users
- Established partnerships to expand AI capabilities and workforce

Key FY27 Deliverables:

- 9 operational AI hubs (initial 6)
  - Deploying 3 FMs from FY26
- 6 domain-specific foundation models trained on initial curated datasets
- 20 DOE AI FM applications developed and deployed
- Suite of curated datasets to enable further model development
- "100 AI EF" systems deployed
- Upgraded compute infrastructure to support model training and inference
  - Supporting 2,000 DOE active scientific/engineering users
- Established partnerships to expand AI capabilities and workforce

Key FY28 Deliverables:

- 12 operational AI hubs
  - Deploying 6 FMs from FY27
- 8 domain-specific foundation models trained on initial curated datasets
- 30 DOE AI FM applications developed and deployed
- Suite of curated datasets to enable further model development
  - Add more science, energy and security topics
- "200 AI EF" systems deployed
- Upgraded compute infrastructure to support model training and inference
  - Supporting 5,000 DOE active scientific/engineering users
- Expanded partnerships to expand AI capabilities and workforce

Key FY29 Deliverables:

- 12 fully operational AI hubs
  - Deploying and supporting 10 world leading FMs from FY28
- 10 updated domain-specific foundation models trained on curated datasets and synthetic data
- 40 DOE AI FM applications developed and deployed
- Suite of curated datasets to enable further model development
  - Partnerships with industry on synthetic data augmentation
- "500 AI EF" systems deployed
- Upgraded compute infrastructure to support model training and inference
  - Supporting 10,000 DOE active scientific/engineering users
- Mature partnerships to sustain AI capabilities and workforce

Key FY30 Deliverables:

- 12 fully operational AI hubs
  - Deploying and supporting 10 world leading FMs from FY29
- 12 updated domain-specific foundation models trained on curated datasets and synthetic data
- 60 DOE AI FM applications developed and deployed
- "1000 AI EF" systems deployed
- Suite of curated datasets to enable further model development
  - Partnerships with industry on synthetic data augmentation
- Upgraded compute infrastructure to support model training and inference
  - Supporting 20,000 DOE active scientific/engineering users
- Mature partnerships to sustain AI capabilities and workforce

The key elements are phased in incrementally each year, with 6 initial AI hubs and 3 foundation models in FY26, growing to 12 hubs and 12 mature FM models by FY30, and 60 FM based AI applications. Investments in computing, data curation, partnerships and other enabling capabilities also scale up year-over-year in proportion to the overall budget growth.