BMS EXPERIENCE WITH NVIDIA SUPERPOD

Bill Mayo, BMS

Senior Vice President for Research Technology

Bristol Myers Squibb

H100 SuperPOD

PRISME Fall 2024

Greg Meyers, EVP, Chief Digital & Technology Officer

"Digital innovation, computer and data science are the disciplines that will create breakthroughs in our understanding of biology and transform the way the life sciences industry operates at a fundamental level,"

"Our goal is to increasingly leverage digital innovations across all aspects of our business, and we can only achieve that by partnering with those who are at the forefront of digital innovation."

Bill Mayo, SVP Research Business Insights & Technology

"Biology is Computation"

Focus on business need not technology

Challenge: Provide scaled compute to enable aggressive use of ML for drug discovery

Consideration	Cloud GPU	SuperPOD	
Deployment Speed	+	- (one time)	High demand limited ability to consume at scale without commitments, eroding normal cloud consumption-only pricing advantage
Scalability		+	High demand changed the normal flexibility and scalability equation for cloud providers
Configurability	+	+	Container model gives full configuration capabilities in either environment
Tooling	-	+	Moving toward neutral as NVIDIA open sources tooling like BioNemo, but initially favorable to purchase
Cost Efficiency	-	+	Extremely high demand, and corresponding prices combined to reverse normal cloud economics
Enable bold compute	-	+	Scientists spending time trying to predict costs or limiting creativity for fear of running up a bill
Encourage exploration	-	+	Scientists spending time trying to predict costs or limiting creativity for fear of running up a bill

Bristol Myers Squibb – SuperPOD – Equinix Private AI - Research AI Platform

Research AI Platform for BMS

(LLMs + GenAI + Transformers)

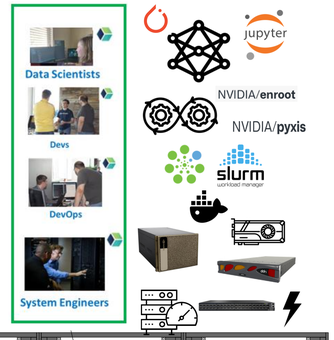
(GNNs + Diffusion Models + HF + Imaging + Protein Modeling + Geometric DL)

Mark III Co-Pilot & User Engagement Services

(Success Support for Research Users & Research IT Ops)

(Assist at 3 layers of the stack & Build custom BMS KnowledgeBase)

(Constant parallel code stack validation in Mark III DGX Pod Lab)



NVIDIA Software

(AI Enterprise + BCM)

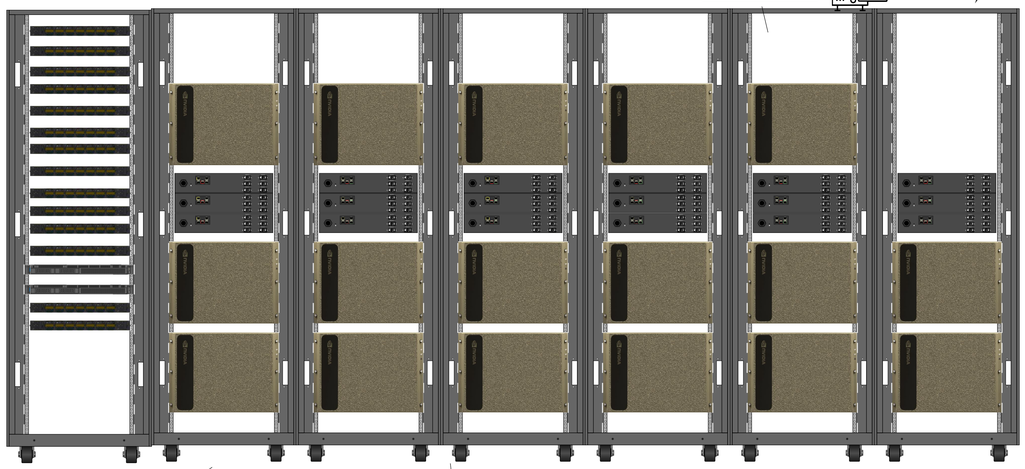
(BCM + Slurm)

Equinix Co-lo Datacenter & Managed Svcs

("Private Cloud" – Ashburn, VA)

DGX Cloud

(Azure – 6 Nodes, 2 months)



Also:
BasePod A100 cluster
100+ A100's
Some single A100's
Cerebras
AWS Bursting
AWS Native

SuperPOD Networking (NVIDIA)

(18x IB NDR + 5x Ethernet)

SuperPOD (NVIDIA DGX H100)

(31x DGX H100 – 1SU)

Mark III Stabilization & Troubleshooting (Pre-SuperPOD)

(6 months continuous prior to SuperPOD on existing BasePOD to ensure future DGX direction)

Teaming with NVIDIA Team

(Strategic/Tactical & Creative Engagement with BMS together with NVIDIA at all levels)

SuperPOD Storage (DDN)

(4PB AI400X2 + 2PB Intelliflash)

AI Education Series (Future)

(Work w/ BMS to draw in non-traditional users for SuperPOD)

Apparently this is what it looks like

2024 SuperPOD Sample Uses

Initiative	Team	Status	Impact
Develop foundation model for molecular glue discovery	Discovery & Development Sciences - NLCC	Complete	Models developed and applied towards library expansion efforts. Prioritized >11,000 compounds for screening
Build conditional generative models for small molecule design	Discovery & Development Sciences - NLCC	Complete	Models developed and applied towards library expansion. Designed novel cores for synthesis by medicinal chemistry.
Fine-tuning of AlphaFold for Antibody-Antigen Complexes	Discovery & Development Sciences - Biotherapeutics	On-going	Cross project support. Biotherapeutics discovery programs
High content phenotypic screening	Discovery & Development Sciences - LDO	On-going	Gain access to more GPUs to scale existing capabilities (challenges transitioning the code from Domino to SuperPOD), AWS Savings
Models for analyzing images from Oncology trials	Global Drug Development - GBDS	On-going	Development of cutting-edge multi-modal foundational models for image analysis for accelerating clinical trials, delivering speed improvement, enabling faster iteration and increased research. AWS savings
Co-folding models in predicting ternary structures with CRBN	IPS - MOCR, Oncogenesis and PH	New	Define SOTA models performance retrospectively on data relevant to the task & provide insights for future modeling
Scaling transformer-based inference for genome-wide perturbation screens	IPS – CICT & Neuro	On-going	Gain access to more GPU compute to scale inference resulting in target identification. Reduce AWS cost.
Assess and deploy Foundation Models for single-cell transcriptomics	IPS - Knowledge Science Research	On-going	In-silico perturbation. Future beneficiaries will be all the stakeholders willing to test multiple models on their own data.
Assess and deploy Foundation Models for histopathology and multiplexed immunofluorescence	IPS – Knowledge Science Research	On-going	Apply foundation models to patient stratification, clinical trial outcome prediction and disease understanding. Current and future beneficiaries of the projects are Translational IPS, Translational Medicine, and Translational Pathology.

Where Next

Short Term

- Simplifying onboarding
- Migrating things that benefit
- Driving demand through engagement, simplification, training, stories, etc.

Long Term

- Assigned 4-year life (v. normal 5) in recognition of pace of change
- NVIDIA DGX Model allows managing DGX Cloud and OP/Colo as single cluster
- We have a couple years to watch the economics develop
- All options open