



Scale Day - Sept. 20 2022

SPECTRUM SCALE (GPFS) AT NYU LANGONE HEALTH

John Speakman, Assistant Vice President, Research and Education IT Michael Costantino, Director, High Performance Computing Paul Glick, HPC Senior System Engineer

NYU Langone Health - Research Data

- cell biology
- radiology
- pathology
- genetics
- computational medicine population health

genomics

biochemistry

neuroscience

cardiology

neurology

cancer research

pulmonary research

medical education

image processing

infectious diseases

endocrinology

nephrology

- microbiology
- radiation oncology
- ophthalmology
- psychiatry
- environmental medicine other
- dermatology
- rheumatology
- autoimmunity
- radiosurgery
- orthopedics
- gastroenterology



These systems support and accelerate the research of a broad range of Bioinformatics, Life Sciences, and Machine Learning.

- ~16PB Research Data
- 200+ Research Labs
- 2000+ Users



Genomics - SARS-CoV-2



Integrated Genome Viewer (<u>https://software.broadinstitute.org</u>) plots of sequencing coverage of the entire length of the spike gene (3,821 nt, x-axis) of 3 representative sequences of the Omicron variant and 3 representative sequences of the Delta variant of severe acute respiratory syndrome coronavirus 2, processed in the same batch.

"Dr. Adriana Heguy, the director of the Genome Technology Center at NYU Langone Health and professor at the Department of Pathology, NYU Grossman School of Medicine, is involved in the genomic surveillance of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Her team has submitted more than 6,000 sequences to GISAID-the database that is the world's largest repository for SARS-CoV-2 genomic sequences. "

Sarah Ahaley, Shital. (2022, February 07). Omicron sequences containing Delta mutation detected during SARS-CoV-2 genomic surveillance. News-Medical. Retrieved on September 20, 2022 from https://www.news-medical.net/news/20220207/Omicron-sequences-containing-Delta-mutation-detected-during-SARS-CoV-2-genomic-surveillance.aspx.



Cell Biology - Cryo-EM

"Potassium is the major intracellular cation in all kingdoms of life. In bacteria, potassium is used to maintain membrane potential, pH, and to provide turgor pressure for cell growth and division. When external potassium levels are low, an ATP-dependent pump called KdpFABC is essential to maintain intracellular potassium levels.We used cryogenic electron microscopy to investigate this unique partnership and to address mechanisms of energy coupling and transport. We present structures corresponding to all major reaction intermediates and map distinct densities along the proposed transport pathway. Based on this data, we describe a mechanism that is unique amongst known membrane transporters."

Sweet ME, Larsen C, Zhang X, Schlame M, Pedersen BP, Stokes DL. Structural basis for potassium transport in prokaryotes by KdpFABC. Proc Natl Acad Sci U S A. 2021 Jul 20;118(29):e2105195118. doi: 10.1073/pnas.2105195118. PMID: 34272288; PMCID: PMC8307911.



Fig. 1. Overview of cryo-EM structures. (A) Cryo-EM density maps of KdpFABC in four unique conformational states, corresponding to intermediates in the Post–Albers reaction cycle. KdpA is green; KdpB is brown, blue, yellow, and red; KdpC is purple; and KdpF is cyan.

Machine Learning - Radiology



A schematic representation of how we formulated breast cancer exam classification as a learning task. The main task that we intend the model to learn is malignant/not malignant classification. The task of benign/not benign classification is used as an auxiliary task regularizing the network.

Deep Neural Networks Improve Radiologists' Performance in Breast Cancer Screening

"We present a deep convolutional neural network for breast cancer screening exam classification, trained, and evaluated on over 200000 exams (over 1000000 images). Our network achieves an AUC of 0.895 in predicting the presence of cancer in the breast, when tested on the screening population."

> N. Wu *et al.*, "Deep Neural Networks Improve Radiologists' Performance in Breast Cancer Screening," in *IEEE Transactions on Medical Imaging*, vol. 39, no. 4, pp. 1184-1194, April 2020, doi: 10.1109/TMI.2019.2945514.



Machine Learning - Neuro surgery



NYU Langone physicians, led by Dr. Eric K. Oermann and Dr. Daniel A. Orringer, continue to advance novel applications of artificial intelligence, with implications across the continuum of medicine.

Background: Extended postoperative hospital stays are associated with numerous clinical risks and increased economic cost. Accurate preoperative prediction of extended length of stay (LOS) can facilitate targeted interventions to mitigate clinical harm and resource utilization.

Objective: To develop a machine learning algorithm aimed at predicting extended LOS after cervical spine surgery on a national level and elucidate drivers of prediction.

Health

Conclusion: Machine learning algorithms accurately predict extended LOS across single-center and national patient cohorts and characterize key preoperative drivers of increased LOS after cervical spine surgery.

Valliani AA, Feng R, Martini ML, Neifer SN, Kim NC, Gal SJ, **Oermann EK**, Caridi JM. APRIL – **Pragmatic Prediction of Discharge Disposition Following Cervical Spine Surgery with Machine** Learning and Validation on a National Scale. *Journal of Neurosurgery* NYU Langone

Pediatric Psychiatry - Neuroimaging

Fetal Hippocampal Connectivity Shows Dissociable Associations with Maternal Cortisol and Self-Reported Distress during Pregnancy

Maternal stress can shape longterm child neurodevelopment beginning in utero. One mechanism by which stress is transmitted from mothers to their offspring is via alterations in maternal cortisol, which can cross the placenta and bind to glucocorticoid receptor-rich regions in the fetal brain, such as the hippocampus.



Associations between maternal distress and cortisol during pregnancy with fetal hippocampal connectivity. (**A**) Greater self-reported maternal distress was related to stronger positive functional coupling between the hippocampus and right posterior parietal association cortex at p < 0.01 (uncorrected), k > 20. (**B**) Higher maternal cortisol output during pregnancy was associated with stronger positive functional coupling between the hippocampus and dorsal anterior cingulate cortex (dACC) and (**C**) left medial prefrontal cortex (mPFC).

Hendrix CL, Srinivasan H, Feliciano I, Carré JM, Thomason ME. Fetal Hippocampal Connectivity Shows Dissociable Associations with Maternal Cortisol and Self-Reported Distress during Pregnancy. *Life*. 2022; 12(7):943. https://doi.org/10.3390/life12070943



NYU Langone GPFS - Landscape



- 268 filesets
- 3 data pool; scratch (SSD), data (HDD), Metadata (SSD)
- GPFS Capacity: 15PB
- Change Rate: 10TB 20TB/day

- Spectrum Protect Backups and HSM
 - TS4500 Local Tape Library (~26PB)
 - TS4500 Remote Tape Library (~20PB)
 - 500TB Replication Cache
 - 4x TSM Clients
 - 4x HSM Clients



NYU Langone Backup - Challenges

Constraints

- 24 Hour Cycle
- 45 Day Retention
- Off-site Copy for DR
- Shared 10Gb WAN
- 66ms Latency
- 12 Tape Drives
- 4 Backup Clients
- 500TB Cache

Required Processes

- Snapshots
- Backups
- Expiration
- Replication
- Tape Migration
- Reclamation



NYU Langone Backup - Scheduler



Due to the high variance in process times, threshold and cron driven schedules were not sufficient to meet the 24 hour backup window.



NYU Langone Backup - Workload Manager





7.25 hours with workload manager

11 MCIT Example is from 1 of 4 nodes with 5 parallel threads, current system executes on 4 client nodes with a total of 20 threads.



NYU Langone GPFS Backup (cont.)

Work load manager output:

Starting the Backup Workload Manager: 2022/09/16 00:00:49

Number of Servers=4 Number of Slots=20 Client Servers: hpc-pr2-proto3.ib => 5 hpc-pr2-proto1.ib => 5 hpc-pr2-proto1.ib => 5 To Be Backed Up: 268

Fri Sep 16 00:00:50 EDT 2022

Polling: hpc-pr2-proto3.ib, Running(1,)Open Slots: 4

bin/launcher.sh backup gpfs data_cryosparc_data data_cryosparc_data_2022-09-16-00-00 bin/launcher.sh backup gpfs data_cryoem_data data_cryoem_data_2022-09-16-00-00 bin/launcher.sh backup gpfs data_sequence data_sequence_2022-09-16-00-00 bin/launcher.sh backup gpfs data_armachelab data_armachelab_2022-09-16-00-00 Polling: hpc-pr2-proto2.ib, Running: 0, Open Slots: 5

bin/launcher.sh backup gpfs data_johnsonplab data_johnsonplab_2022-09-16-00-00 bin/launcher.sh backup gpfs data_aifantislab data_aifantislab_2022-09-16-00-00 bin/launcher.sh backup gpfs data_carrolllab data_carrolllab_2022-09-16-00-00 bin/launcher.sh backup gpfs data_kirchhofflab data_kirchhofflab_2022-09-16-00-00 bin/launcher.sh backup gpfs data_HernandoLab data_HernandoLab_2022-09-16-00-00 Polling: hpc-pr2-proto1.ib, Running: 0, Open Slots: 5

bin/launcher.sh backup gpfs data_coudraylab data_coudraylab_2022-09-16-00-00 bin/launcher.sh backup gpfs data_giannarellilab data_giannarellilab_2022-09-16-00-00 bin/launcher.sh backup gpfs data_evronylab data_evronylab_2022-09-16-00-00 bin/launcher.sh backup gpfs data_devinskylab data_devinskylab_2022-09-16-00-00 bin/launcher.sh backup gpfs data_bhabhaekiertlabs data_bhabhaekiertlabs 2022-09-16-00-00

Polling: hpc-pr2-proto4.ib, Running: 0, Open Slots: 5

bin/launcher.sh backup gpfs home home_2022-09-16-00-00 bin/launcher.sh backup gpfs data_tsirigoslab data_tsirigoslab_2022-09-16-00-00 bin/launcher.sh backup gpfs data_molecpathlab data_molecpathlab_2022-09-16-00-00 bin/launcher.sh backup gpfs data_proteomics data_proteomics_2022-09-16-00-00 bin/launcher.sh backup gpfs data_luilab data_luilab_2022-09-16-00-00 Currently Running: 20 Remaining: 249

Fri Sep 16 04:19:16 EDT 2022

Polling: hpc-pr2-proto3.ib, Running(1) Open Slots: 4 Polling: hpc-pr2-proto2.ib, Running: 0, Open Slots: 5 Polling: hpc-pr2-proto1.ib, Running: 0, Open Slots: 5 Polling: hpc-pr2-proto4.ib, Running: 0, Open Slots: 5 Currently Running: 1 Remaining: 0

Fri Sep 16 04:21:17 EDT 2022

Polling: hpc-pr2-proto3.ib, Running: 0, Open Slots: 5 Polling: hpc-pr2-proto2.ib, Running: 0, Open Slots: 5 Polling: hpc-pr2-proto1.ib, Running: 0, Open Slots: 5 Polling: hpc-pr2-proto4.ib, Running: 0, Open Slots: 5 Currently Running: 0 Remaining: 0

Backup Workload Manager Complete: 2022/09/16 04:21:18







QUESTIONS?





THANK YOU

