

ComPrePS 2.0: Enabling Massive-Scale Distributed Computing on High-Performance Computing Cluster for Histopathological Data Processing

Suhas Katari Chaluva Kumar^{1, 3}, Anindya S. Paul¹, Haitham Abdelazim¹, Will Dunklin², David Manthey², Oleksandr Moskalenko⁴, Sayat Mimar¹, Erik Deumens⁴, and Pinaki Sarder^{1, †}

¹Dept. of Medicine – Section of Quantitative Health, University of Florida, Gainesville, FL

²Kitware, Inc, Clifton Park, NY

³Dynanet Corp., Elkridge, MD

⁴Research Computing, Information Technology, University of Florida, Gainesville, FL

ABSTRACT

System automation and artificial intelligence (AI) have significantly impacted the healthcare sector, leading to the rapid digitization of histological data into gigapixel resolution whole-slide images (WSI). This advancement has fueled the emergence of computational pathology, necessitating high-resolution image processing on a massive scale. To meet this need, a distributed computing environment with extreme scalability and secure architecture is essential for biomedical workloads involving large amounts of private data..

At Computational Microscopic Imaging (CMI) Laboratory, we developed and open-sourced the Computational Renal Pathology Suite (ComPrePS). This image analysis tool integrates histopathological slide visualization, quality evaluation, and AI-driven data analysis for clinical tasks. ComPrePS has already been utilized by the Kidney Precision Medicine Project (KPMP) to process tissue slides and generate data crucial for understanding kidney disease progression.

The initial implementation, ComPrePS 1.0, offered a user-friendly interface and AI-driven analysis but faced limitations due to its monolithic architecture, inadequate scalability, and security concerns. These challenges prompted the development of ComPrePS 2.0, which integrates unparalleled computing power from high-performance computing clusters (HPCs) using University of Florida's (UF) HiperGator—one of the world's largest supercomputers. By leveraging on-demand CPU, GPU and MEM resources, Apptainer-based containerization, and parallel file system access, ComPrePS 2.0 achieves unprecedented scalability and improved security. Benchmarking on 920 large-dimensional WSIs demonstrated a 15× performance improvement over its predecessor. ComPrePS 2.0 aims to facilitate truly large-scale automated histopathological slide analysis, benefiting researchers, pathologists, and students worldwide.

Keywords: High performance computing (HPC), AI/ML Image Analysis, Scalable AI systems, Distributed computing, GPU-accelerated compute, Histopathology Analysis, HiperGator.

1. INTRODUCTION

We are at a transformative moment in the field of medicine. Advancements in Imaging technologies has led to rapid digitization of histological data, generating big data pools of whole slide images (WSI) from human biopsies. At the same time, recent breakthroughs in AI, demonstrated by innovations like ChatGPT and NVIDIA's GPU revolution have unleashed unprecedented possibilities across industries, with healthcare at the forefront. In pathology, AI is revolutionizing traditional workflows that historically depended on painstaking manual analysis by pathologists. Tasks once performed under a microscope are now streamlined through computational pathology, where AI analyzes gigapixel WSIs with precision and at scale.

Institutions like the National Institutes of Health (NIH) have already recognized this potential with initiatives such as The Human BioMolecular Atlas Program (HuBMAP) (1) and Kidney Precision Medicine Project

†Send correspondence to Pinaki Sarder, PhD E-mail: pinaki.sarder@medicine@ufl.edu

(KPMP) (2). These projects provide substantial funding to foster collaborative research environments aimed at developing novel computational tools in advancing our understanding of human body and disease types at the cellular level. Another significant contemporary development is HiperGator SuperComputer by University of Florida (UF). UF has partnered with two leading edge hardware providers, NVIDIA (3) and AMD, to develop a one of the world's most powerful super-computing infrastructure that provides a highly scalable and distributed computing environment capable of meeting the extensive computational demands of modern biomedical image analysis workloads.

This confluence of advances led to the development of the Computational Renal Pathology Suite (ComPrePS). ComPrePS (4) is a cloud-based platform designed for the analysis, visualization, and organization of histopathology whole slide images (WSIs) based on Digital Slide Archive (DSA)(5). Key features of ComPrePS include efficient data management and storage, interoperability with access to several histological image formats such as .tiff, .czi, .svs, .ome-tiff and results generated in various user-friendly formats such as JSON, XLSX, CSV, and XML. The browser based platform also automates the segmentation of kidney micro-anatomical structures such as glomeruli, tubules, arteries/arterioles, and peritubular capillaries incorporating state-of-the-art (SOTA) AI algorithms for complex image analysis. Furthermore, ComPrePS is an open-source, easy-to-use software, accessible to users with little to no AI or software coding expertise. (See (6) for video demonstration of ComPrePS.)

While ComPrePS 1.0 (7) was met with success due to its user-friendly interface and one-button image analysis on WSIs, it faced limitations in performance, scalability, and security that inhibited its massive potential. Specifically, it was unable to access the massive hyperscale computing environment such as HiperGator because it was not designed for this level of integration. A major architectural redesign of ComPrePS was thus crucial to harness the capabilities of the HiperGator.

2. METHODOLOGY

The original architecture of ComPrePS 1.0 was monolithic (see Fig. 1) and consisted of five core building blocks: Girder, Worker, Slicer-CLI, HistomicsUI, and Mongo database (DB)(7). Girder handles the data and user management and is extensible for deploying any user-developed AI models, thus making the system extremely powerful. The MongoDB facilitates rapid operations on various histological image formats (such as .tiff, .svs, .czi). HistomicsUI offers a user-friendly interface for visualization, annotation, and metadata addition. The Worker component handles heavy AI/ML workloads, while Slicer-CLI manages AI plugins and job monitoring.

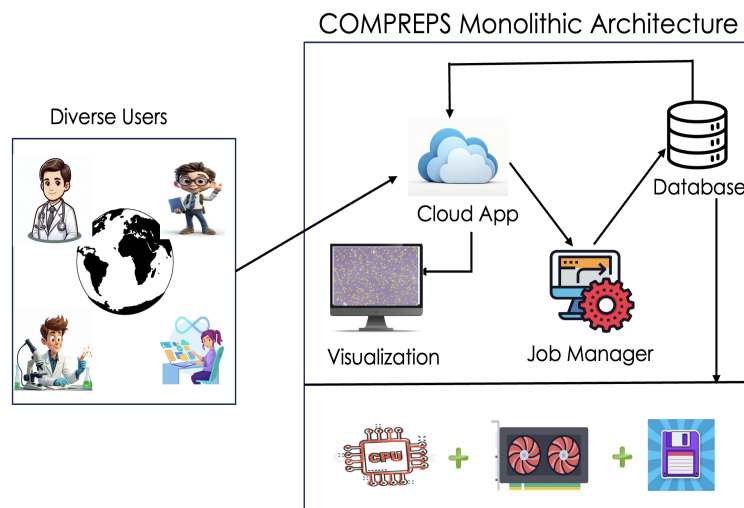


Figure 1: ComPrePS 1.0 - Monolithic Architecture.

ComPrePS 1.0 was designed with a Docker-based architecture. However, Docker containers pose security concerns for deploying in a multi-user computing environment such as HPCC as docker has an inherent

tendency to run with superuser privileges. Docker is also incompatible with resource scheduling systems like Simple Linux Utility for Resource Management (SLURM), which is commonly used in High performance Computing Clusters (HPCC) to manage multi-user workload requirements. Consequently, the system was constrained to computing resources within a single node, limiting scalability and preventing access to HPCC resources.

In order to address the issues of security and compatibility with HPCC systems, transitioning to Apptainer containerization became essential. Apptainer is compatible with HPCC environments, offering enhanced security, fine-grained control over resource scheduling, and efficient utilization of computing resources. Unlike Docker, Apptainer operates with minimal privileges and performs well in HPCC systems, as demonstrated in the results section.

The new re-engineered architecture involved significant changes in two core components: Worker and Slicer-CLI. These components were decoupled from the Docker-based architecture and transitioned to a Apptainer-based, SLURM-compatible distributed architecture. In this new setup, the main component, Girder, still resides on a single, minimal-resource node; However, the computationally intensive image analysis workloads are handled by the Worker, utilizing on-demand, dynamically allocated computing nodes on the HPCC system. All storage needs are met by HiperGator's Lustre-based parallel file system, which can be scaled rapidly. These compute nodes accessing high-speed parallel file system, can run image analysis and AI jobs with significantly higher resources than what was possible with ComPrePS 1.0, significantly enhancing performance and scalability while also ensuring efficient usage of large datasets without data duplication.

ComPrePS 2.0, with its new architecture (see Fig. 2), now offers a powerful, scalable, and secure platform for researchers, pathologists, and students. The system has unlimited potential, constrained only by the resource investments made in terms of hardware resources such as CPUs and GPUs.

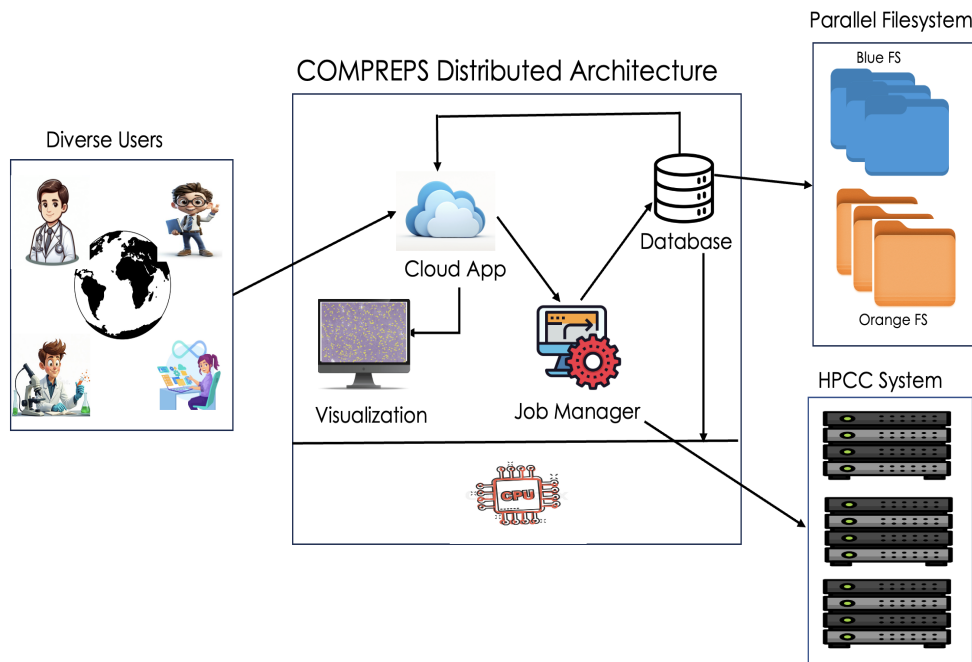


Figure 2: ComPrePS 2.0 - Distributed Architecture with HiperGator Integration.

3. EXPERIMENTS AND RESULTS

Benchmark Dataset

The experiment dataset consisted of kidney biopsies from 920 patients and one tissue section from each patient. The average size and dimension of each WSI is approximately 350 MB and 100K×50K squared pixels. The average number of micro-anatomical structures such as glomeruli is around 30 and tubules is around 1500 per WSI. Each RGB pixel in WSI is saved in *uint8* datatype, so total dataset size is 322,000,000,000 Bytes.

Task

We have chosen our custom image segmentation pipeline, multi-compartment segmentation (8), to demonstrate the improvement of slide processing time of ComPrePS 2.0 compared to the previous version. The details of kidney specific multi-compartment segmentation pipeline and performance metrics can be found at (9).

Results

As shown in Fig 1, ComPrePS 1.0 was deployed on a single node featuring 28 Intel Xeon CPU @ 2 GHz, 2 NVIDIA Quadro RTX 6000 GPU cards, and 8 TB of storage. The CPU cores, based on an older x86 architecture and operating at lower clock speeds, were not optimized for resource intensive workloads such as AI. Similarly, the GPUs were less powerful and comprised of significantly less CUDA cores and vRAM compared to modern variants like the NVIDIA A100 or H100 GPUs, accessible via HPG. Initially, the system had 2 TB of storage, which had to be scaled up to 5 TB and later to 8 TB, necessitating continual expansions of the single-node storage capacity.

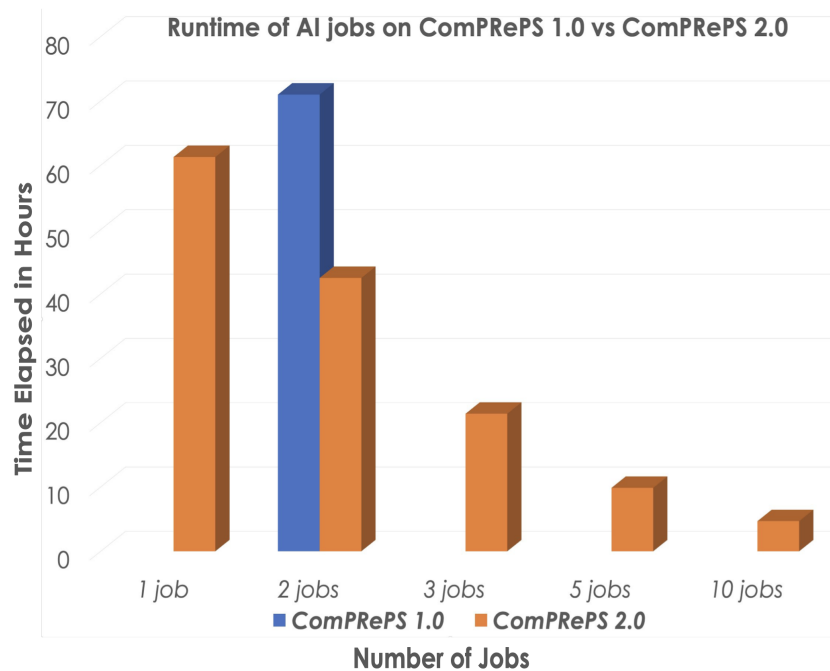


Figure 3: Performance comparison between ComPrePS 1.0 and ComPrePS 2.0

The ComPrePS 2.0 architecture's main app is hosted on a dedicated node with 8 AMD 32-Core CPUs @ 3 GHz, providing markedly improved computational power. As shown in Fig 2, AI workloads referred as "jobs" were executed using dedicated compute node at runtime. Each node can leverage NVIDIA A100 GPU with 80 GB vRAM. This node also has access to the parallel file system, which means our application has access to high speed theoretically unlimited file storage, facilitating easy and faster access to data without data duplication.

Performance Comparison

As illustrated in Fig. 3, the performance benchmarks between ComPrePS 1.0 and ComPrePS 2.0 revealed substantial improvements. Only one data point was available for ComPrePS 1.0 due to its use of just 2 GPUs for the GPU-centric workloads (The number of GPUs were not expandable due to the constraint of available GPU slots in the deployed node). The runtime for the dataset using our benchmark dataset and task was 71 hours for ComPrePS 1.0. In contrast, ComPrePS 2.0 completed the same workload in 42.5 hours with a single GPU, significantly outperforming its predecessor. Moreover, with 2 GPUs, ComPrePS 2.0 reduced the runtime further to 21.4 hours. In addition, ComPrePS 2.0 had access to 10 GPUs and it

completed our task in just 4.7 hours demonstrating at least 15X performance improvement. This exemplifies the exceptional computational capabilities and efficiency of the upgraded architecture.

4. CONCLUSION AND FUTURE WORK

In this paper, we introduce the next iteration of ComPrePS, documenting significant improvements in performance and scalability achieved through leveraging the HiperGator Supercomputer. These advancements enhance the capacity to execute complex and sophisticated AI/ML image analysis models, aligning with our overarching mission to democratize AI in digital pathology and facilitate its widespread adoption by clinicians, students, and researchers, all while achieving markedly reduced processing times and necessitating no coding experience.

Demonstrating significant potential, our system's performance is now primarily constrained by our investments in research computing infrastructure. We have ambitious plans for further enhancement, guided by the Findable, Accessible, Interoperable, and Reusable (FAIR) (10) data principles established by the NIH. Our commitment to transparency and collaboration will be exemplified through the integration of Globus and InCommon federated authentication systems. Globus enables secure and reliable sharing of large datasets across institutional boundaries, while InCommon federated authentication facilitates streamlined access to our application via HiperGator federated login.

A distinguishing feature of ComPrePS is its extensive repository of AI models referred to as the 'Model Zoo'. We continuously strive to expand this repository by adding new AI models tailored to suit the diverse needs of our users. Concurrently, we are also dedicated to optimizing existing models to maximize performance and further reduce runtimes.

Our mission is to engage a broader audience of pathologists, researchers, and students, encouraging them to leverage our open-source systems to accelerate their research. Therefore, we invite interested individuals to learn more about our system and get started with ComPrePS by visiting our [Wiki Page](#) (4).

5. ACKNOWLEDGEMENTS

The authors would like to acknowledge the support from the National Institute of Health (NIH) / National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) grants R01 DK114485, R21 DK128668, U01 DK133090 and A Computational Image Analysis Platform (CIMAP) for the Human BioMolecular Atlas Program (HuBMAP) grant OT2 OD033753.

REFERENCES

- [1] Jain, S. et al., "Advances and prospects for the human biomolecular atlas program (hubmap)," *Nature Cell Biology* **25**, 1089–1100 (July 2023).
- [2] de Boer et al., "Rationale and design of the kidney precision medicine project," *Kidney international* **99**(3), 498–510 (2021).
- [3] "UF Announces \$70 Million Artificial Intelligence Partnership with NVIDIA — news.ufl.edu." <https://news.ufl.edu/2020/07/nvidia-partnership/>. [Written 21-07-2020].
- [4] "ComPrePS | Revolutionizing Digital Pathology." <https://compreps.wiki/>.
- [5] Gutman, D. A. et al., "The Digital Slide Archive: A Software Platform for Management, Integration, and Analysis of Histology for Cancer Research," *Cancer Research* **77**, e75–e78 (10 2017).
- [6] CMILab, "ComPrePS | Watch how it works." <https://rb.gy/pkp98p> (2024).
- [7] Mimar, S. et al., "Compreps: an automated cloud-based image analysis tool to democratize ai in digital pathology," in [*Medical Imaging 2024: Digital and Computational Pathology*], **12933**, 245–251, SPIE (2024).

- [8] SarderLab, “Sarderlab/multi-compartment-segmentation.” <https://github.com/SarderLab/Multi-Compartment-Segmentation>.
- [9] Lucarelli, N. et al., “Correlating deep learning-based automated reference kidney histomorphometry with patient demographics and creatinine,” *Kidney360* **4**, 1726–1737 (Nov. 2023).
- [10] Wilkinson, M. D. et al., “Addendum: The fair guiding principles for scientific data management and stewardship,” *Scientific Data* **6** (Mar. 2019).