Dmitry Volkov

cell: 415 964 8807 | email: dsvolk@gmail.com | Redwood City, CA 94065 | https://www.linkedin.com/in/dsvolk/
Master's degree in experimental physics

PROFESSIONAL SUMMARY

I have 15+ years of experience in system optimization and machine learning, with a strong track record of success (company awards, promotion from regional office to HQ). I have over 7 years of experience leading teams in organizations such as Salesforce, ServiceNow, and currently at Cruise, along with direct management experience at MariaDB. I have successfully managed projects across a diverse range of hardware and software products (backend systems, SaaS offerings, Cloud offerings), ensuring each release meets exceptional performance standards. My key strengths include identifying areas for improvement, setting clear objectives, and driving results.

I enjoy explaining technical concepts to non-experts and can communicate effectively with both technical and non-technical audiences. I simplify complex topics and find solutions that everyone agrees on. I focus on innovative problem-solving and have been awarded a U.S. patent, with a second patent currently in progress. I'm a fast learner, which allows me to quickly align with a company's priorities.
I've worked on projects ranging from creating performance automation for stress testing (Xbench, MariaDB) to developing simulation software for upcoming migrations (MariaDB to Postgres mass migration, ServiceNow) and modeling in resource-constrained environments (Customer Performance Experience, Salesforce).

My Open-Source contribution:
AI Performance notes: https://aiperf.run
Github: https://github.com/the-dsvolk

Recent Education:
- Leadership Principles for Software Engineers, Coursera
- Nvidia AI Infrastructure and Operations Fundamentals, Certificate
- Parallel Programming with CUDA using C++, Johns Hopkins, Coursera
- Kubernetes and Docker from Linux Foundation

SKILLS

- Leadership
- Capacity Planning and Optimization
- Strategic Thinking
- Project Management
- Problem-solving
- Cross-Functional Collaboration
- Linux OS/kernel tuning and optimization
- Docker and Kubernetes
- Splunk, Prometheus, Grafana
- Jenkins,
  Clouds: AWS, GCP, Azure
- Python, Bash, SQL
- Nvidia GPU, InfiniBand, 100+ GbE, RoCE, CUDA, NCCL
- Performance Analysis and Troubleshooting
- Release performance verification
- Google SRE 4 Golden Signals
- USE (Brendan Gregg)
- Performance experiments
- Software Optimization
- Developing performance tools
- CPU multi-threading, memory, and IO optimization.

EXPERIENCE

Cruise, Remote, CA - Capacity and Performance Engineering, Machine Learning Systems
November 2024 - Current

H100 efficiency program: I optimized resource allocation for NVIDIA H100 Kubernetes clusters to enhance performance and efficiency of large-scale AI workloads. I improved GPU utilization and reduced job queueing time through workload placement and scheduling strategies by 20%. I used NVIDIA Nsight System(nsys), Nsight Compute (ncu) and PyTorch profilers to analyze and resolve pipeline bottlenecks and Intel Perfstats to optimize data loading. I leveraged Google BigQuery, Looker, Chronoshphere to monitor system performance and guide data-driven optimization efforts. Worked with internal teams to set up standards for future ML model performance (using NVIDIA sm_occupancy and sm_active metrics). Tools used: Nvidia DCGM metric exporter, nvidia-smi(performance counter from the GPU hardware), nvtop. Tuned NCCL

communication between PODs (Drive 575xx, CUDA 12.9). Optimized workload by NUMA affinity to Pytorch training.
Future-looking GPU system/evergreen strategy: I managed an external contractors team to evaluate GPU server SKUs, memory configurations, and CPU-GPU ratios to improve cost-performance for ML training and inference workloads. In order to achieve this I've built a performance lab (Kubernetes, Helm, Docker, Terraform) to test H100, H200 and B200 in the Google Cloud. My work allowed the team to stay within a budget of 100M per year while keeping the SLA of 2.5B ticks per year.
Fine-Tuning Hyperscaler infra: I optimized low level aspects of containerized workload leveraging of deep understanding compatibility between GCP Container-Optimized OS, Nvidia driver, CUDA and NCCL versions, GCP Infiniband plugin. It was shown that for multi-node workloads performance improvements is about 10% (in terms of rows per second).

ServiceNow, Santa Clara, CA — Senior Staff Capacity Engineering
May 2023 - Nov 2024

Top Customer List Assessment: I collaborated with the SWAT team to build a unified top customer list. Each team had its own evaluation criteria for identifying top customers, so this was a cross-team effort to establish a single source of truth. I used a weighted metrics approach, assigning different weights to database metrics (CPU usage, memory usage, disk) and business metrics. I used Impala to extract raw data (SQL queries) and performed data analysis using Jupyter Notebooks and Python Pandas. This work helped to save 5M in infrastructure cost by providing the right type of resources.
AI Search Capacity Model: I evaluated the existing capacity model for the AI Search application, which had run out of capacity. I used XGBoost, t-SNE, and K-means clustering to build a new model. AI Search Java-based applications generated a large number of metrics, and my task was to identify the most relevant ones. I demonstrated that some critical features had been overlooked, and incorporating them improved the model's accuracy and reliability. This project saved about 20M USD for one year.
Database Migration Modeling Tool: I developed a modeling tool to assist with the mass migration from MariaDB to PostgreSQL. Once all MariaDB instances are migrated, the existing servers can be redeployed to host PostgreSQL databases. This problem is a variation of the multidimensional bin packing problem. Due to power and space constraints in our data centers, doubling capacity during migration was not an option. My model carefully predicted the optimal sequence in which databases should be moved to minimize the number of new servers required while staying on schedule. This helped to save 15M in infrastructure cost.

MariaDB, Redwood City, CA — Director, Performance Engineering
May 2021 - Apr 2023

Management experience: I led a team of five engineers testing various companies' products, including distributed MySQL-compatible databases such as Xpand and MariaDB. I established team processes for gathering and publishing data in a unified format. I grew the team through direct hiring, internal transfers and contracts. I also made some difficult decisions regarding team composition, including letting a few people go. I've managed a budget of 10M per year for Cloud.

Technical leadership: I optimized the performance of the company's cloud offerings. I worked closely with development teams, suggesting which features they had to implement to be competitive in the cloud. My responsibilities include release testing, production testing, and nightly build testing. In addition, I did competitive testing, which includes PostgreSQL, AWS Aurora, Cockroach DB, GCP AlloyDB. My daily activities included reading code base (written in C), debugging Kubernetes performance in AWS, testing hugepages for various databases, and testing different CPUs (Intel, AWS, Graviton 2/3) for database workloads. My work supported SkySQL - company's Cloud offering (ARR of $47.4 million with 650+ customers)

With my team, we open-sourced our benchmarking tool: **https://github.com/mariadb-corporation/xbench-community**

Salesforce, San Francisco, CA — Principal Engineer, Infrastructure Performance and Capacity
Dec 2014 - Apr 2021

Production incident analysis: As a member of the Infrastructure and Performance team, I took on the challenge of solving the quiz: Was the recent production incident a performance or capacity issue? Given the scale of Salesforce's thousands of production machines, this problem couldn't be solved manually. I developed models to describe system utilization based on incoming traffic. These models captured data from tests to predict the future capacity requirements after the new release was deployed.

Customer Performance Experience: I developed a new approach to the measured availability of the company systems. From a customer perspective, if the system is not performing well, it is unavailable. My objective was to determine what 'well' meant in a multi-tenant environment where every customer had its unique implementation. I developed the Customer Performance Experience (CPE) model. My model allowed us to accurately calculate true system availability, a value confirmed by subject matter experts.

Job scheduling optimization: I worked on algorithms to optimize the scheduling of customer jobs, improving efficiency and resource utilization for the Customer 360 engineering team. This involved developing strategies to handle job prioritization, load balancing, and execution timing. The environment included Kubernetes for container orchestration and Spark for large-scale data processing. My work allowed the team to satisfy the SLA of processing 10B records per hour.

EDUCATION

National Research Nuclear University “MEPhI”
Moscow, Russia — *Master's degree/Magister in Physical Sciences*