White Paper

Enterprise Application Workload Optimization

Autonomous Runtime Optimization of Containerized Applications with Granulate on Red Hat® OpenShift®

Real-time Continuous Optimization Solution by Granulate, an Intel company, optimizes workloads to reduce compute spending by up to 60% and improve key performance metrics by up to 40%. Certification of Granulate for Red Hat® OpenShift® Container Platform validates stability and provides for coordinated support, building further on customer value for cloud-native deployments.

Table of Contents

The overhead costs associated with running large-scale workloads can be significant enough to impact business competitiveness and profitability. Organizations therefore have a strategic imperative to optimize software to deliver on business requirements while consuming the least amount of compute resources possible. Such tuning can be costly and tends to distract technical teams from other work. It frequently requires code to be rewritten to achieve performance goals. Moreover, static code optimizations depend on the OS scheduler to allocate resources, which cannot adapt to changing usage patterns and data flows.

Real-time Continuous Optimization Solution by Granulate, an Intel company, overcomes these obstacles for application workloads that operate on thousands of processor cores simultaneously. It allows organizations to handle compute workloads with 60% fewer servers while improving performance by 40%, with no code changes required. Granulate reads into and learns the specific patterns of resource usage and the data flow on a per-application basis, as shown in Figure 1.

This framework enables Granulate to identify contended resources, bottlenecks and prioritization issues. Based on that understanding of the workflow and its requirements, Granulate formulates OS-level scheduling and prioritization

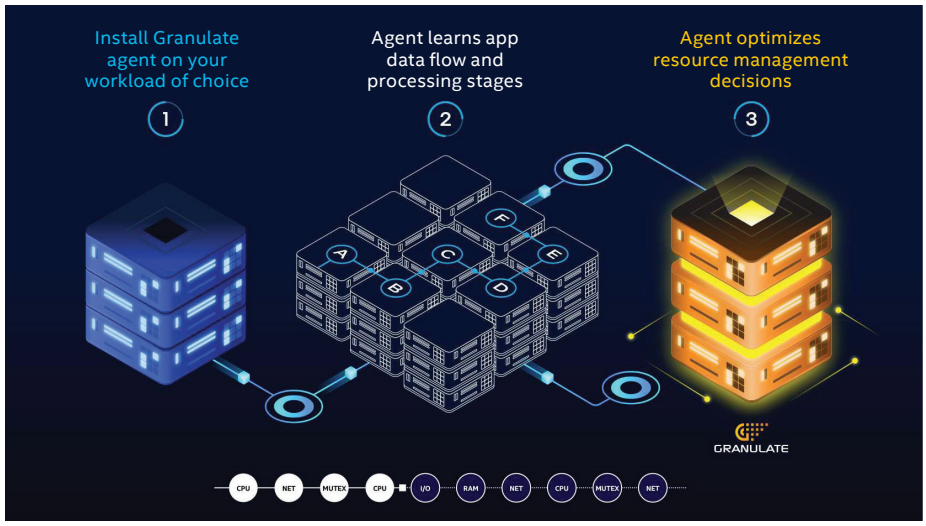


Figure 1. Automatic identification and resolution of bottlenecks and resource contention within application dataflows.

guidance that it maintains and updates in real time. Improved data flow and resource utilization dramatically increases application performance without code changes and reduces costs by up to 50% or more. To help advance the cloud-centered enterprise, Granulate is tuned and certified for deployment on Red Hat OpenShift. This engineering partnership provides assurances of stability and quality to customers as they implement Granulate in their container environments.

Real-Time Continuous Optimization with Granulate

Granulate provides visibility into workflows to improve dramatically on the general-purpose resource scheduling provided by server operating systems. To enable real-time continuous optimization, Granulate automatically learns an application's specific resource usage patterns and data flow to identify contended resources, bottlenecks and prioritization opportunities in each server. It then tailors OS-level scheduling and prioritization through decisions regarding compute resources, locks, caches and memory accesses to improve the infrastructure's application specific performance and enable significant cost reduction. Granulate has three primary components:

- **Granulate gProfiler** continuously profiles code, visualizing application execution sequences and resource usage with a powerful, intuitive user interface. Comprehensive code profiling helps development teams identify sections of code that are good candidates for optimization, to reduce compute requirements and costs.
- **Granulate gCenter** is a dashboard used to monitor, visualize and track Granulate's benefit to the organization using key performance and cost metrics, as shown in Figure 2. It provides summary statistics and enables drilling down into specific clusters or services to identify additional savings opportunities.
- **Granulate gAgent** is deployed on hosts to provide visibility into workflows and resource usage and allocation patterns. It supports on-premises or SaaS deployments, on bare-metal, in virtual machines (VMs) or in containers. The gAgent can be deployed using a command-line interface (CLI) , using deployment software such as Red Hat Ansible, or using a Kubernetes DaemonSet.

The application-driven optimization performed by Granulate includes custom thread scheduling according to real-time processing requirements. The network stack operates in lockless fashion, helping take maximum advantage of massive parallelism to enable high throughput and resource efficiency. Granulate provides efficient inter-process communication using modern protocols and shared memory, avoiding excess overhead. Intelligent connection pooling reduces overhead further, without requiring application changes, autonomously providing congestion control among those connections. Granulate also tailors memory allocations and accesses based on the outcomes of usage-pattern analysis.

Real-World Results from Diverse Environments

Real-world businesses across verticals have realized significant improvements across metrics and key performance indicators (KPIs) that include costs, latency, throughput and CPU utilization. These benefits accrue across a broad spectrum of technology platforms, such as Java, Python, Go and Big Data, as presented in Figure 3. Granulate customer use cases demonstrate the ability to realize these benefits with rapid time-to-value and little friction, in as little as two weeks. No customization of existing code is required.

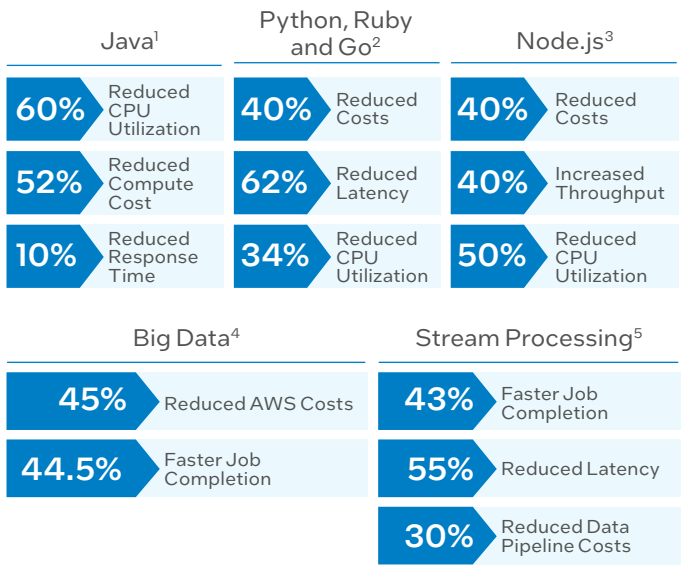


Figure 3. Granulate customer proof points.

Figure 2. The gCenter Overview dashboard.

Flexible Support Across Large-Scale Compute Requirements

As a fundamental enabler of real-time workflow optimization, Granulate flexibly supports a range of large-scale computing needs, as illustrated in Figure 4. It is designed for Linux architectures, which run the vast majority of enterprise production workloads, especially those that operate across thousands of processor cores. Specifically, Granulate targets customer infrastructures of 15,000 cores or more, with at least one single workload that uses more than 1,000 cores on its own. Those processor resources may be on-premises, cloud, multi-cloud or hybrid, aligning use of the technology with evolving infrastructures used by today's enterprises.

Granulate features installation with a single command line on many different platforms, with support for private, multi-cloud and hybrid-cloud infrastructures. The installation can also be incorporated seamlessly with existing infrastructure-as-code workflows.

As IT organizations advance along their cloud transformation journeys, Granulate is ready to support their changing needs, including containers, microservices and certification for use with Red Hat OpenShift. The software roadmap supports emerging requirements and helps customers on their ways to adopting new approaches such as DevOps, DevSecOps and continuous integration/ continuous deployment (CI/CD). It is also a prime enabler for strategic initiatives such as infrastructure reduction and automation that help apply new technologies so businesses can scale intelligently with positive impacts on data center requirements, headcount and software licensing.

Business and Technology Benefits

Deployment of Granulate is easy, fast and automated, and it can take advantage of any provisioning tool in use. The implementation process does not require any engineering effort by the customer organization, with no administrative overhead. It delivers immediate, real-world benefits, as illustrated in Figure 5.

Cloud-Native Operation with Red Hat OpenShift

Granulate is enabling implementations with cloud-native containerized environments and applications through certification for use with Red Hat OpenShift Container Platform. OpenShift hardens and adds functionality on top of Kubernetes that makes it truly enterprise-ready, as illustrated in Figure 6. The platform is developed using an open-source model to harness community innovation and to enable the ongoing evolution of the container ecosystem. It also integrates directly and completely with the rest of the Red Hat tools portfolio for optimized efficiency.

Certification on OpenShift helps ensure the most robust implementation possible of Granulate for real-time continuous optimization of cloud workloads. The certification process provides for coordinated support and proactive cross-validation for emerging workloads and product releases to protect the customer experience. Certification documents the successful completion of a rigorous array of tests that verify stability and performance under the full spectrum of possible network conditions. This process helps reduce risk and accelerates time-to-value.

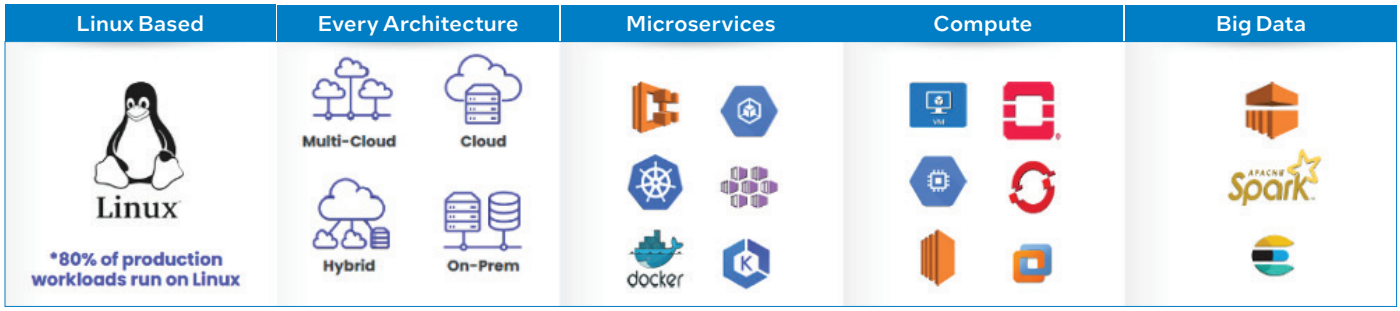


Figure 4. Broad environment support.

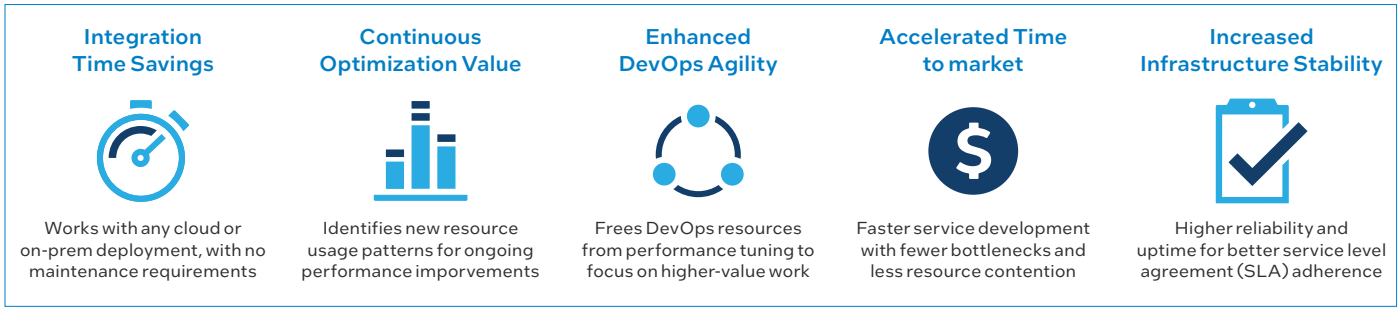
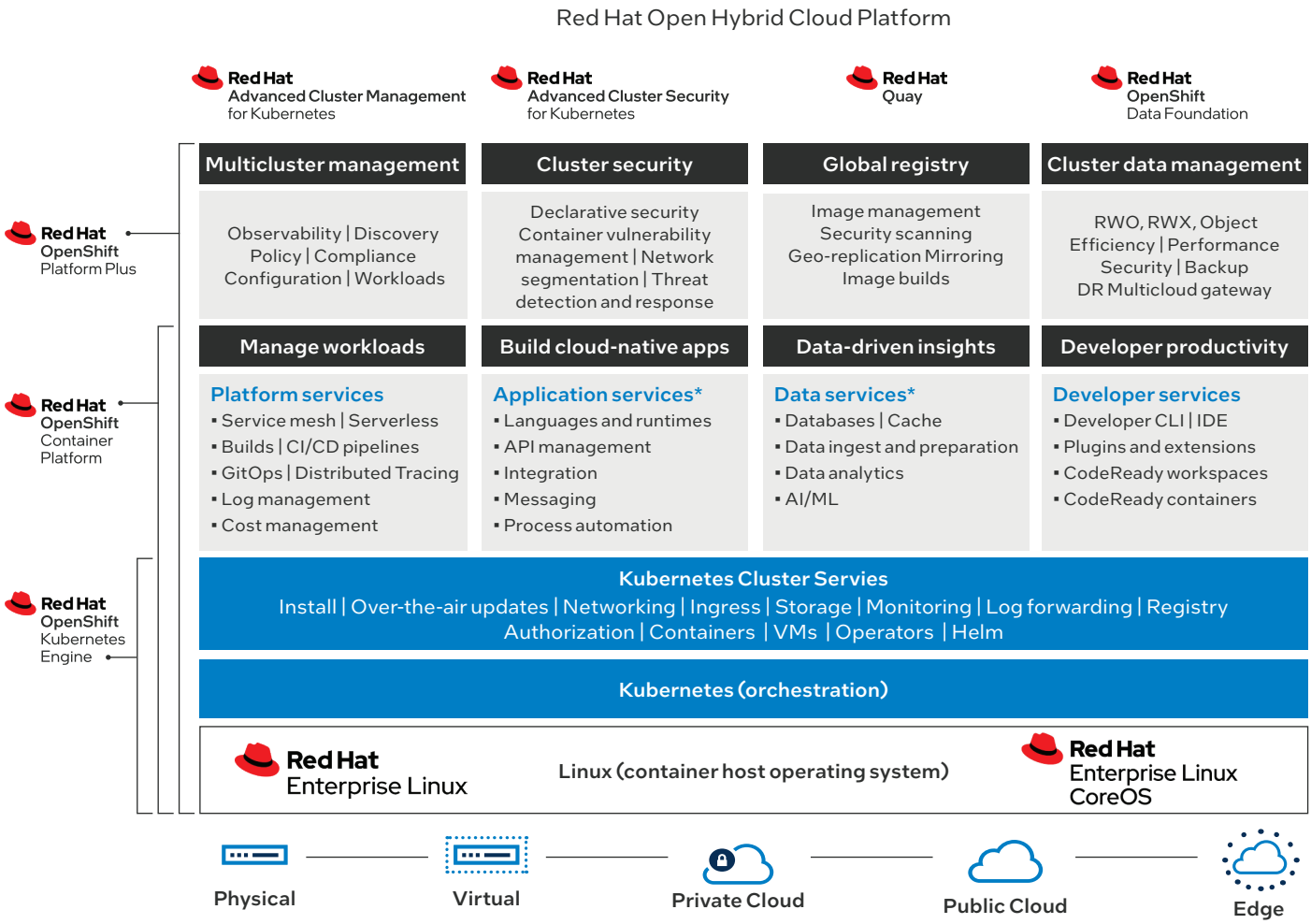


Figure 5. Real-world benefits.



* Red Hat OpenShift® includes supported runtimes for popular languages/frameworks/databases. Additional capabilities listed are from the Red Hat Application Services and Red Hat Data Services portfolios.

** Disaster recovery, volume and multicloud encryption, key management service, and support for multiple clusters and off-cluster workloads requires OpenShift Data Foundation Advanced

Figure 6. Red Hat OpenShift provides additional hardening and capabilities to augment Kubernetes.

The OpenShift ecosystem of solutions brings together self-service efficiency and flexibility for development teams, even as it scales their operations to new heights as new requirements emerge. Streamlining the software lifecycle at any scale compounds the benefit of real-time continuous optimization based on Granulate, taking optimal advantage of Intel architecture features and capabilities. Fast, consistent workload handling responds to fluctuating demand reliably and flexibly, delivering high responsiveness and excellent customer experience while helping IT organizations keep their eyes solidly on the bottom line.

Cluster, platform, application and developer services provide a compelling evolutionary path for forward-looking container environments. Granulate has plug-and-play support for customers running applications on OpenShift, regardless of the cloud provider they use. Instead of deploying Granulate profiling tools or agents using Kubernetes DaemonSet, Granulate can deploy agents in real-time with OpenShift operators.

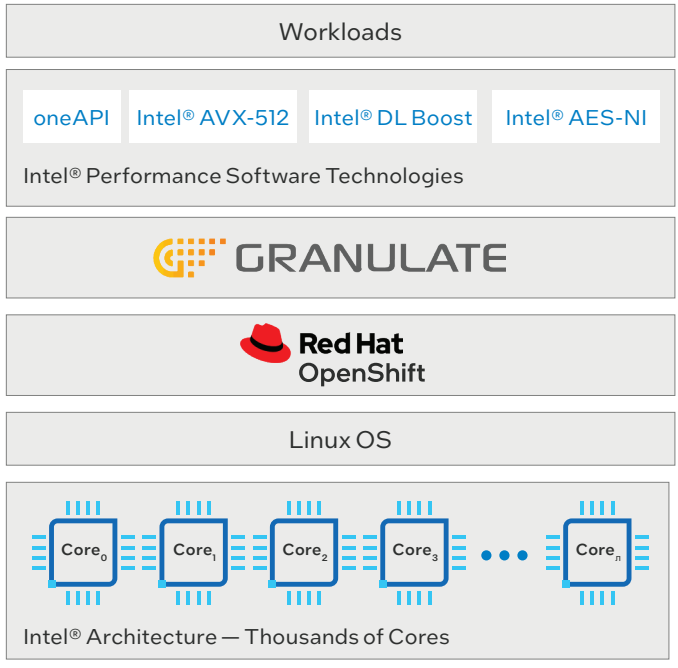


Figure 7. Hardware and software building blocks.

Red Hat OpenShift can be deployed as an on-premises infrastructure, with Red Hat OpenShift Container Platform (OCP), or via cloud managed service offerings such as Red Hat OpenShift on AWS (ROSA), Azure Red Hat OpenShift (ARO), IBM Cloud, and several others. The Granulate solution is supported on all of these infrastructure solutions, as well as on Red Hat OpenShift Data Science (RHODS), which is Red Hat's AI and data science platform.

Intel Hardware and Software Building Blocks

Granulate embraces the holistic set of Intel architecture and other building blocks, which include an extensive software ecosystem as well as hardware advances, as illustrated in Figure 7. Granulate therefore positions customers to realize benefits across their operations, with full fidelity across the entire universe of present and future Intel innovations.

Open Development Across Hardware Architectures

A range of Intel oneAPI toolkits support the programming specification with best-in-class compilers, performance libraries, frameworks and tools for analysis and debugging. Key components used in the context of Granulate include the following:

- **Intel oneAPI Video Processing Library (oneVPL)** is a performance library to optimize media transcode performance across integrated and discrete GPUs with a single codebase. oneVPL provides a video-focused API for fast video decoding, encoding and processing.
- **Intel oneAPI Compilers** generate optimized code that supports industry standards while taking advantage of built-in technology for Intel processors and accelerators. They seamlessly integrate dynamic recompilation at runtime to generate code that is tailored to the execution environment.
- **Intel oneAPI Threading Building Blocks (oneTBB)** is a flexible performance library that enables shared-memory parallelism to improve multicore performance, without specialized multi-threading expertise. Rather than requiring developers to work directly with threads, the library enables developers to specify tasks that are then mapped to threads.
- **Intel oneAPI Math Kernel Library (oneMKL)** is a set of highly optimized and parallelized math routines that support capabilities such as linear algebra routines, random-number generators, vector operations and Fast Fourier Transforms (FFTs). Standard oneMKL computations can be run on Intel GPUs using OpenMP.

Performance Features of Intel Hardware

Running on thousands of Intel cores, Granulate optimizes large-scale workloads during execution to take advantage of present Intel microarchitectural features, with a shared roadmap looking ahead to future silicon generations. Intel Xeon Scalable processors feature high per-core performance to accelerate throughput and increase density in the data center. They incorporate hardware workload acceleration for deep learning/AI, encryption processing, compression and other workloads.

Increased Throughput with Ultra-Wide Vector Operations: Intel AVX-512

To accelerate data-intensive calculations, Intel AVX-512 instructions enable workloads to optimize the amount of data they can operate on per processor clock cycle in Intel Xeon processors. Intel AVX-512 provides ultra-wide 512-bit vectors, into which applications can pack as many as 32 double precision or 64 single precision floating point operations per clock cycle. The technology provides a software instruction set that can be used to accelerate scientific simulations, financial analytics, artificial intelligence (AI)/deep learning, 3D modeling, media processing, cryptography and data compression.

Enhanced Efficiency for Deep Learning/AI: Intel DL Boost

Helping run complex deep learning tasks such as model training, image classification, speech recognition, language translation and object detection on the same hardware as other workloads, Intel Deep Learning Boost (Intel DL Boost) accelerates those computations without requiring the use of an add-on hardware accelerator. Based on Intel AVX-512, Intel DL Boost Vector Neural Network Instructions (VNNI) combine three instructions into one, improving utilization of compute and cache resources. Intel DL Boost also supports the brain floating-point (bfloat16) number-encoding format, which improves efficiency for workloads that have high computational requirements without requiring high precision.

Hardware Acceleration for Encryption: Intel AES-NI

Advanced Encryption Standard (AES) is the most widely deployed encryption standard for protecting network traffic, sensitive data and corporate IT infrastructures. With increased uptake of cloud computing, where data and code leave traditional protected IT settings, encryption plays an increased role in the enterprise. Intel AES New Instructions (Intel AES-NI) is a set of seven software instructions that implement processing-intensive parts of the AES algorithm in hardware, dramatically accelerating cryptographic workloads while increasing security by protecting against side-channel attacks on AES.

Conclusion

Granulate enables enterprises to improve their quality of service while also saving on the capital expenditures (CapEx) and operating expenses (OpEx) that would be associated with deploying more infrastructure. It is available as a certified application for OpenShift in the Red Hat Ecosystem Catalog. Granulate aligns resource allocation with dynamic service data flows that vary over time during execution. Rather than relying on OS-level resource allocation — which was designed for general-purpose use — Granulate tailors allocation to specific application data flows and processing stages. This capability is especially critical for applications that operate across thousands of cores.

Real-time, continuous workflow optimization with Granulate takes advantage of both Intel software and hardware innovations. The oneAPI programming model and associated tools enhance throughput, reduce latency and increase cost efficiency across hardware for CPUs, GPUs and other accelerators. Workloads also benefit from key hardware features that include Intel AVX-512, Intel DL Boost and Intel AES-NI.

To support and advance cloud-native initiatives in customer environments, Granulate is certified for use with Red Hat OpenShift Container Platform. Dynamic workloads supported by containerized microservices can therefore attain a higher level of performance than would otherwise be possible. These advances enable enterprises to reduce the infrastructure burden associated with supporting critical applications by as much as 50% or more, doing more with less and preparing for expansion into new services and opportunities.

More Information

Real-time Continuous Optimization Solution by Granulate, an Intel company: https://granulate.io/product/

Granulate Certification for Red Hat OpenShift (Red Hat Ecosystem Catalog): https://catalog.redhat.com/software/container-stacks/detail/5ed63fe4b906fc8b7bdf2dd3

Red Hat OpenShift: https://www.redhat.com/en/technologies/cloud-computing/openshift

Intel Development Tools: https://www.intel.com/content/www/us/en/developer/tools/overview.html

intel + Red Hat + GRANULATE

[1] Granulate, "How Perion cut compute costs by 52% with no R&D efforts," https://granulate.io/case-studies/perion/.

[2] Granulate, "How MoEngage Achieved 40% Cost Reductions on AWS With Granulate," https://granulate.io/case-studies/moengage/.

[3] Granulate, "Dream11 Improves Kafka Workload Performance and Reduces AWS Costs by 40%," https://granulate.io/case-studies/dream11/.

[4] Granulate, "Mobileye Reduced 45% On Their AWS Costs Leveraging Granulate," https://granulate.io/case-studies/moblieye/.

[5] Granulate, "How Granulate helps Singular reduce cost by 35% on mission-critical services on Amazon ECS," https://granulate.io/case-studies/singular/.

0822/RKM/MESH/350490-001US