

Optimize Application-level Performance, Cut Costs and Boost Efficiency

RESEARCHED BY

OMDIA

COMMISSIONED BY

intel

GRANULATE
An Intel Company

Introduction

Intel® Granulate™ is designed to optimize the performance of applications and workloads by dynamically adjusting the CPU's resource allocation in real time. It analyzes the behavior of the application and the system it runs on. Based on this analysis, Intel Granulate optimizes the CPU's usage, allocating resources more efficiently to improve performance and reduce processing time.

The goal of Intel Granulate is to make applications run faster and more efficiently without the need for code changes or manual optimization. By using this technology, companies can potentially reduce their infrastructure costs, improve user experience, and increase overall system efficiency. This is especially true for complex compute environments such as Kubernetes, compute environments that are allocated to big data analytics, rightsizing infrastructure provisioning on Linux machines and any CPU brand.

WHO SHOULD READ THIS REPORT:

Anyone responsible for provisioning compute infrastructure on-premises or in the cloud. In particular, technical staff including site reliability engineers (SREs), platform engineers, DevOps, big data engineers, Kubernetes engineers, operations engineers, and management staff from FinOps to IT managers, IT directors, and CTOs.



Executive Summary

IT staff often work blind with tools that lack fine-tuning capabilities, resorting to static settings for diverse workloads. Manual optimization can take months for just one variable. Intel Granulate offers an autonomous, dynamic optimization tool for CPU cores (regardless of brand) and rightsizing compute infrastructure, alleviating manual headaches.

In this report we identify three major use case types:

- 1 CPU performance optimization:** Reduced thread locking, minimized latencies, and optimized memory allocation.
- 2 Smart capacity management:** Wastage identification, dynamic provisioning, and rightsizing, unlocking cost savings with the potential for substantial benefits on Kubernetes environments.
- 3 Dynamic optimization of big data analytics:** Optimize costly analytics and data streaming environments by maximizing resources.

The ultimate benefit of Intel Granulate is cost savings by reducing idle and overprovisioned resources. Any data center or cloud operation can benefit, certainly for large enterprise operations at Internet scale where reductions are identified, it equates to massive cost savings in millions of dollars. The alternative of hiring kernel engineers to perform manual optimization is costly, time-consuming, and lacks the ability to dynamically allocate resources across multiple performance related variables. Moreover, such manual fixes are one-time and do not account for feature updates and code changes, whereas Intel Granulate acts continuously.

Intel Granulate's technology is only applicable to Linux OS machines, as it needs to dive deep into the system kernel, and this is not possible with a closed OS. Note however, that Intel Granulate makes no changes on the kernel-level or the OS level, only on the runtime-level and application-level, and no changes are required to code. It does apply to any Linux distribution and is CPU brand agnostic, as well as cloud service provider agnostic. Intel Granulate also complies with the highest security standards with SOC2, ISO, HIPAA, and GDPR certifications.

Executive Summary

The key messages of this report are:

1 Efficient Resource Allocation: Lack of transparency into CPU performance results in cautious over-specification of the number of CPU cores resulting in idle compute waste and overprovisioning. With better insights into CPU performance, organizations can accurately determine the actual resource needs of their applications and workloads. Intel Granulate captures this information and autonomously and continuously applies the optimizations with no R&D efforts required. This leads to more efficient resource allocation, as organizations can provision the right number of CPU cores for each task without overestimating or underestimating requirements. This typically translates to significant cost savings.

Transparent CPU performance metrics allow organizations to identify and address application bottlenecks. They can optimize code and configurations to achieve better performance, ensuring smooth and responsive user experiences.

2 Adaptive Provisioning: There is a general fear of changing the original developer settings or default provisioning specifications, due to not understanding the impact this will have. Overcoming this fear empowers organizations to dynamically adjust provisioning settings based on actual workload demands. Intel Granulate's adaptive provisioning ensures that

resources are aligned with real-time needs, avoiding both underprovisioning (which could lead to performance degradation) and overprovisioning (which incurs unnecessary costs).

By revisiting and updating provisioning specifications as needed, Intel Granulate helps organizations foster a culture of continuous improvement. As applications and workloads evolve, so should their provisioning settings to remain efficient and effective.

3 Cost savings and scalability: Big data environments are expensive, and their provisioning has a significant impact. Big data environments often involve vast amounts of data and resource-intensive processing. Optimizing provisioning ensures that the right number of resources is allocated to each task, reducing overall infrastructure costs.

With optimal provisioning, big data environments can scale more effectively. They can handle increasing workloads and datasets without incurring the overhead of overprovisioning, making them more agile and responsive to business needs.

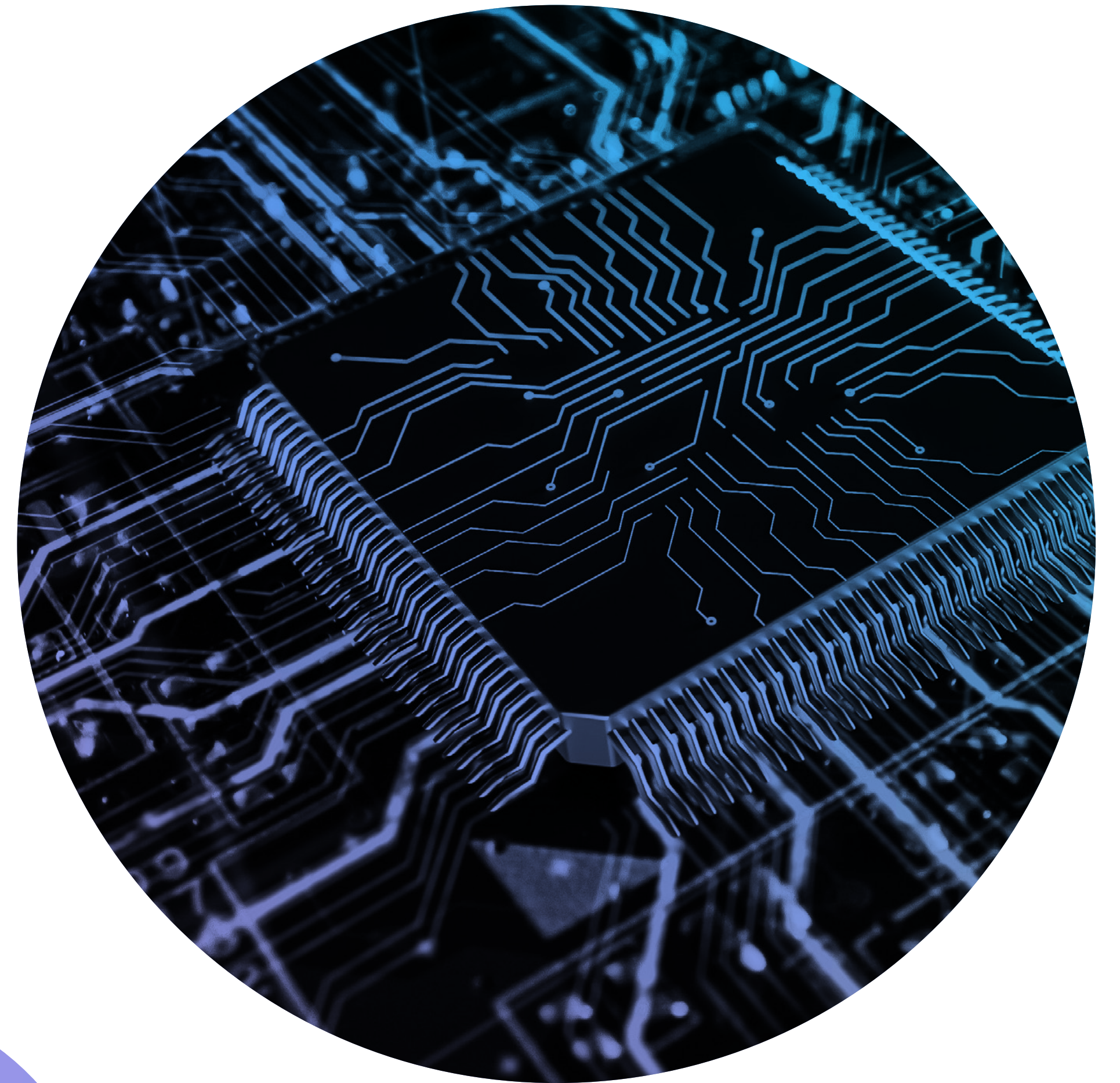
Properly provisioned big data environments operate at their peak performance, leading to faster data processing, analytics, and insights. This increased efficiency can positively impact decision-making processes and business outcomes.

Omdia View

CPU PERFORMANCE TRANSPARENCY EQUALS MORE EFFICIENT RESOURCE ALLOCATION

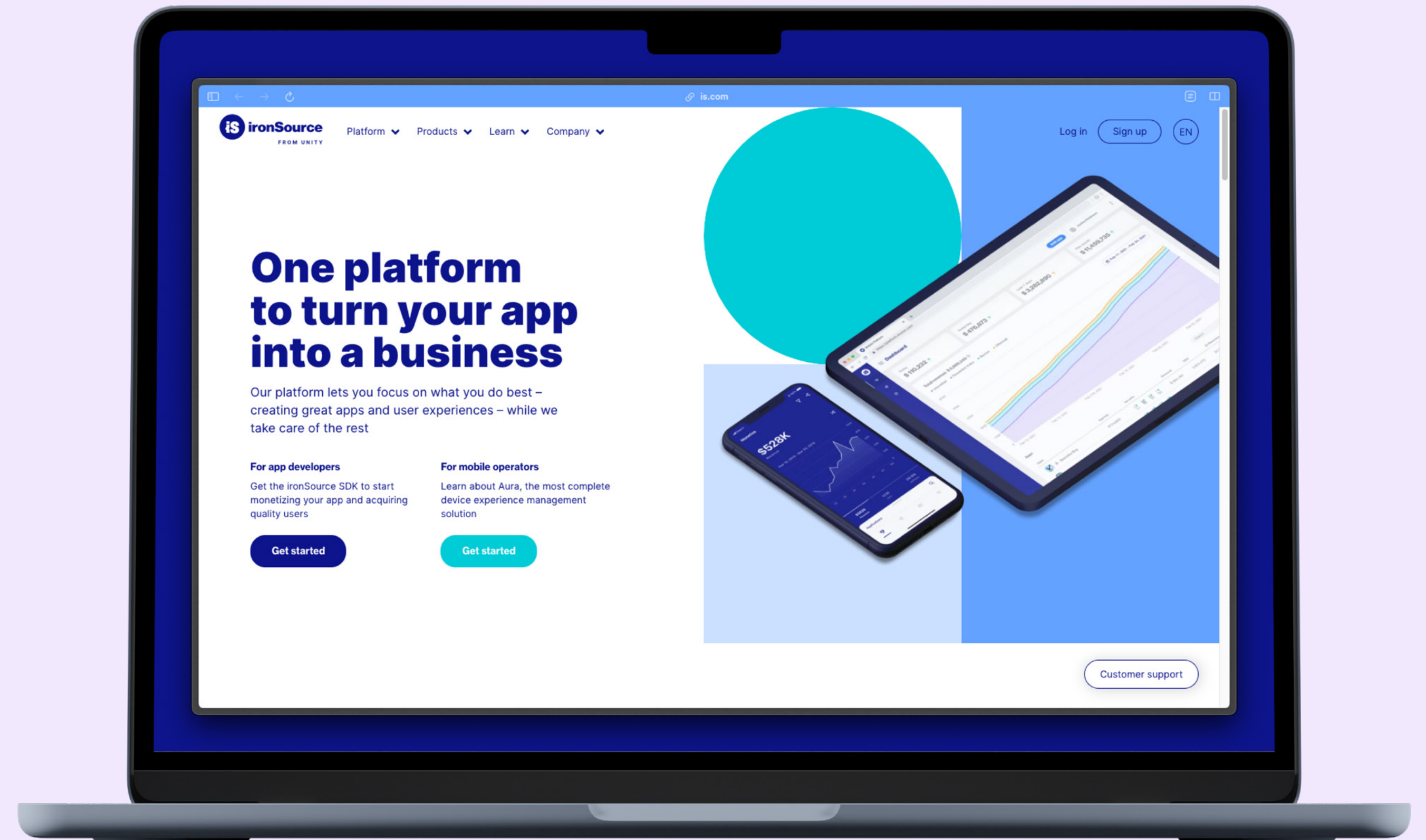
Large organizations perform infrastructure sizing to buffer against burst demands and unexpected contingencies. Cautious provisioning staff typically overprovision. Managed service providers with strict SLA guarantees often find themselves overprovisioned, indeed Intel Granulate sees over 50% of Kubernetes environments being overprovisioned. The biggest contributing factor for this is the lack of transparency into the actual consumption of resources.

Intel Granulate alleviates this lack of visibility by measuring actual resource consumption. With this knowledge, it can confidently set an optimal 20% level of resource overprovisioning, the recommended amount, and do this dynamically and autonomously. The 20% buffer adapts to changing resource consumption, ensuring contingencies are covered. For the right set of applications and at scale, this can translate into huge cost savings.





ironSource is an AdTech software company that turns apps into scalable businesses. Its infrastructure comprises Scala, Node.js on AWS EC2 and EKS. Running 100% on AWS, it makes use of 2k spot instances and five Kubernetes clusters. ironSource receives 25bn daily ad requests, through a 100 TB data pipeline, and uses online and offline machine learning modules processing 4bn daily events. With bidding involved, maintaining low latency is paramount. To test Intel Granulate, the company deployed the Intel Granulate agent on its heaviest, most in-demand servers: the application bidding server. A week after deployment, there was a 20% decrease in instance counts with no change in latency. With this success, the agent was deployed across the entire infrastructure. **Overall, there was a 29% increase in throughput, with a 21% reduction in EC2 instances and a 25% reduction in cloud costs.**



Disclaimer: Results are based on customer analysis performed in 2022. Your results may vary. Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

KUBERNETES RIGHTSIZING

Developers turn to Kubernetes for their cloud-native computing applications, running microservices in containers and using Kubernetes to orchestrate these containers. Kubernetes creates a hierarchy of collections of containers, from pods to nodes and clusters.

Optimizing Kubernetes environments is challenging for the infrastructure engineers who not only need to provision sufficient compute resources for optimal runtime, but also provision Kubernetes at the pod and node levels. Organizations moving to cloud native computing either use default definitions or just focus on using Kubernetes's tools: Cluster Autoscaler which scales nodes and Horizontal Pod Autoscaler (HPA), automatically updates a deployment of applications in a cluster or a set of pods. However, these tools do not have fine-tuning capabilities and over time, overprovisioning accumulation occurs.

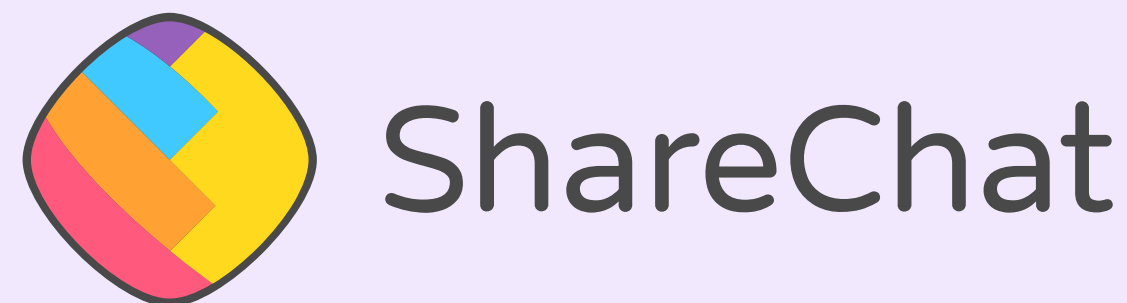
Furthermore, teams developing a product will specify Kubernetes requirements and the provisioning engineers tend not to touch these decisions. This can cause overprovisioning over time as these patterns result in unused resource accumulation. Even teams highly efficient and experienced in using Kubernetes are faced with this challenge.

Working blindly, the engineers overprovision, and many organizations simply live with this type of inefficiency. It is almost impossible to get this data without an inspection tool, and manually configuring this dynamic response across all infrastructure is impossible. Without the data or capability to trust a rightsizing capability while maintaining SLAs, latency, and response requirements, large organizations are intentionally overprovisioning.

Intel Granulate offers a free Capacity Optimization solution, a continuous workload and a pod rightsizing tool for Kubernetes, which eliminates overprovisioning of the Kubernetes environment and thereby reduces costs. The Capacity Optimization solution is designed to provision above what is actually required, allowing for unusual issues or spikes.



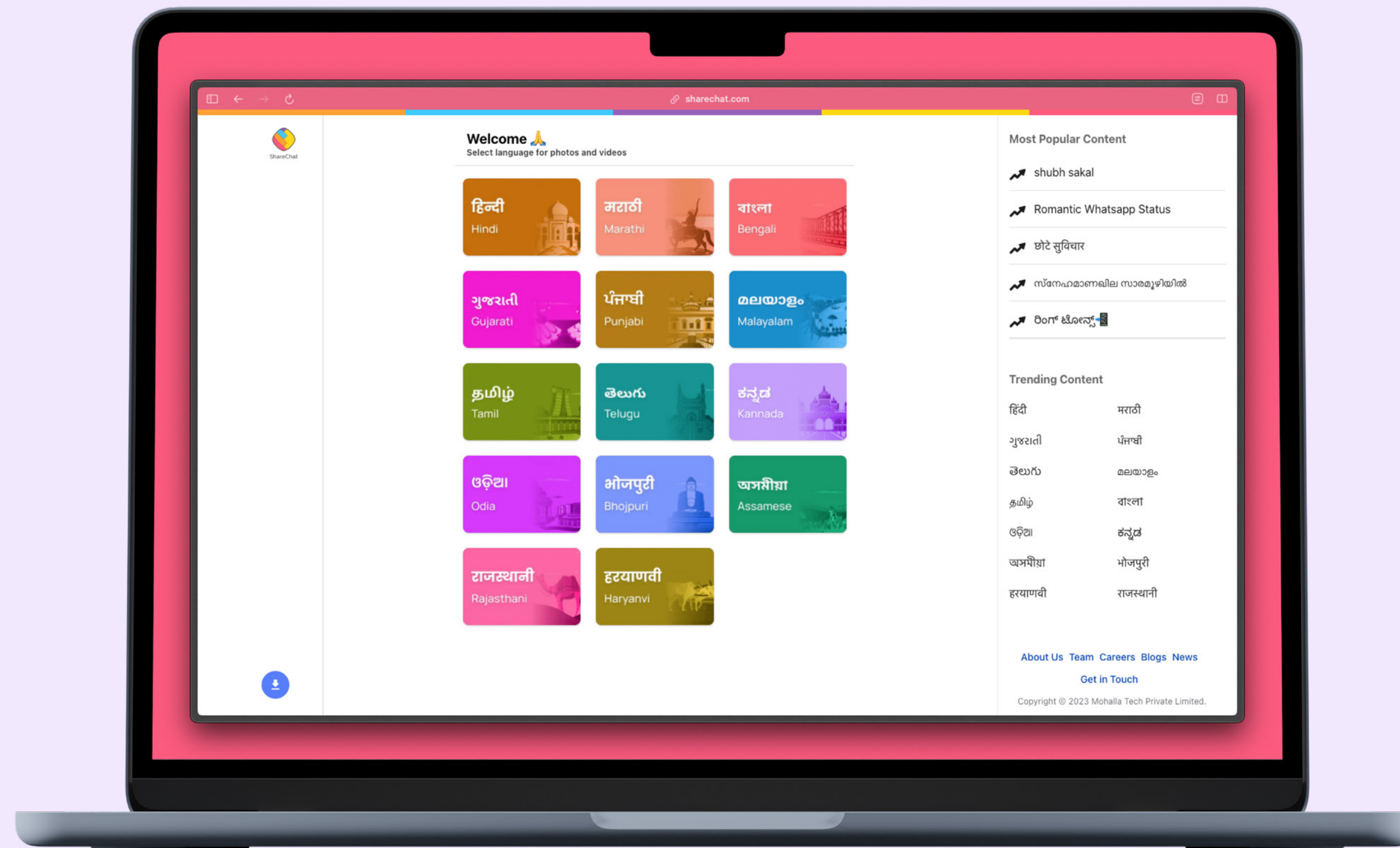
USE CASE:



ShareChat is a social media platform and owner of Moj, the short video app. As a fast-growing platform, it experiences 200k requests per second, nearly 17bn requests a day. The platform runs on Google Cloud and uses Google Kubernetes Engine (GKE), deployed on 50 clusters with over 200k cores. Hands-on tuning did not solve their Kubernetes overprovisioning, and as they scaled out even this manual activity was untenable.

ShareChat turned to Intel Granulate. Starting with orchestration of Kubernetes, they fit resources to actual usage, autonomously right sizing their GKE workloads to avoid both overprovisioning and under-utilization. With the Intel Granulate capacity management solution, they could scale vertically while also using HPA for pod optimization.

Overall, ShareChat achieved a 25% response time reduction, a 25% reduction in CPU usage, and a 20% reduction in compute costs.



Disclaimer: Results are based on customer analysis performed in 2022. Your results may vary. Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

BIG DATA ANALYTICS OPTIMIZATION

In the big data world, organizations tackle batch processing, stream processing, ad-hoc analysis, processing AI/ML input data, and more. But here's the challenge: According to Intel Granulate's estimates, these environments can be incredibly costly: typically a Databricks setup costs 12 times more than a regular VM, and Amazon EMR costs four times as much. Efficiency is paramount to success in this domain.

Processing big data is no small task, involving multiple dimensions like cleansing, re-formatting, and masking for compliance. Tools like Apache Spark and Apache Hadoop are powerful but tough to optimize manually. Without visibility into bottlenecks, data processing speed suffers. As data scales up, so do inefficiencies in unoptimized computing. Rapid changes in scale can further impact efficiency, and even small workload changes affect Kubernetes cluster performance, requiring code and configuration retuning. It's a complex challenge.

Intel Granulate sees data engineering teams having a "set it and forget it" approach for workload configuration. As pipelines, volumes, and sources change, the configuration changes need to be done manually. This is laborious, doesn't scale, and isn't optimal: to optimize performance, the engineers need access to real-time monitoring, logging data pipeline events, and collecting, storing, and visualizing metrics such as job completion time, CPU, and memory usage. They need to assess whether infrastructure resources such as compute, storage, and networking are meeting demand as changes to data occur, and processing needs scale up.

Intel Granulate offers an autonomous solution for these big data challenges that operates continuously, optimizing workloads efficiently across resources. Intel Granulate is infrastructure agnostic and able to optimize on all of the most popular execution engines, like Kafka, Spark, Tez, and MapReduce, platforms like Dataproc, Amazon EMR, HDInsight, Cloudera and Databricks, and resource orchestrations like YARN, Kubernetes, and Mesos. As data volume, variety, and velocity fluctuate, the Intel Granulate agent is constantly updating to ensure that resources are allocated efficiently, with minimal CPU and memory wasted.



USE CASE:

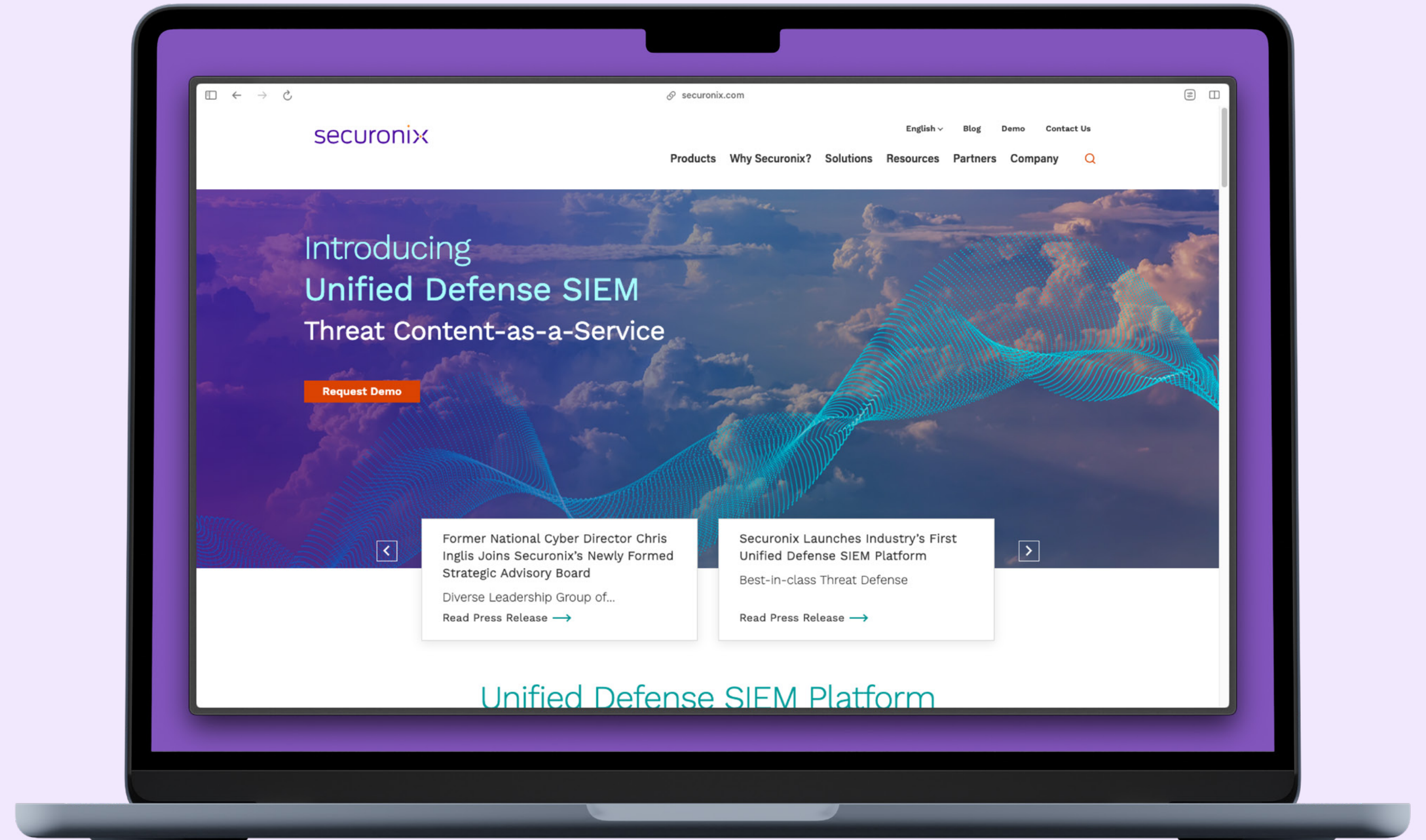


securonix

Securonix is a cybersecurity company offering a security information and event management solution. The company makes use of Apache Spark on Amazon EMR, collecting, analyzing, and responding to massive volumes of data in real time, consequently, most of their operating expenses are attributed to cloud spend, 50% of their overall infrastructure cost was allocated to Spark Streaming on Amazon EMR.

Starting with a proof of concept, Securonix chose five EMR clusters on which to apply Intel Granulate. The result was processing higher amounts of traffic while using fewer cores, and this benchmark met Securonix's 15% infrastructure reduction goal with up to 33% fewer cores in one environment. Intel Granulate automatically and dynamically allocated Spark executor resources based on job patterns and predictive idle heuristics. Intel Granulate's solution also applied continuous YARN optimizations by more efficiently allocating resources based on CPU and memory utilization. With this success, Securonix expanded Intel Granulate fleet-wide and within a few weeks had optimized all 45 EMR clusters. **Deploying Intel Granulate achieved a 15% reduction in infrastructure and 33% fewer cores in one environment, with significant cost reductions.**

Disclaimer: Results are based on customer analysis performed in 2022. Your results may vary. Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.



Technology Overview

OPTIMIZING ANY CPU RUNNING ON LINUX SYSTEMS

Intel Granulate observes how an application runs and optimizes the allocation of the compute resources it consumes. Furthermore, the user does not need to make any changes to the application code. The compute variables that it optimizes are:

- **Latency** – decreases ⬇️
- **CPU utilization** – reduces ⬇️
- **Requests per sec (RPS) and throughput** – increases ⬆️

Intel Granulate performs profile guided optimization, in contrast with static program analysis, Intel Granulate profiles dynamic tests or production runs of the application. Moreover, Intel Granulate continuously optimizes resource allocation, dynamically adjusting resources, unlike tools that fix configurations.

LEARNING PROCESS

Intel Granulate first has to learn the behavior of an application by observing it over a period of one or two weeks. The key requirement is that the entire workload is running to capture the application behavior, the degree of load (number of instances of the application running) is not important. The Intel Granulate profiler, which is open source and free, identifies the runtimes present in the environment and an Intel Granulate agent and its runtime module then hooks onto resource allocation mechanisms and processes at the runtime level of the machine.

For example, in a Java application, the agent will connect with the JVM, hooking into decision making queues such as threads scheduling and runtime resources: it monitors how many resources each thread received, and how long it took to complete a specific function. It improves hot-path function access and performs pre-allocation and release of memory space and object sizes to reduce allocation overhead.

OPTIMIZATION PROCESS

After the learning phase, the agent has identified all the different patterns per thread, it can then actively re-prioritize the allocation of resources between threads. This improves performance and removes bottlenecks. For example, in JVM thread scheduling a mutex lock limits access to a resource by threads, preventing overlapping. Intel Granulate's agent can learn the repetitive patterns of mutex allocation and re-prioritize a mutex lock to a different thread and then re-assign it to that thread when it's more efficient to do so, and by this process, the agent is able to solve mutex bottlenecks.

Another example is in-memory allocation. Processes are assigned a certain amount of memory by the OS, and this initial assignment is inaccurate, so the OS needs to correct itself, and adds more and more memory over the runtime of the process for it to complete its function. These correction iterations result in a higher

consumption of CPU compute. The agent can learn the repetitive patterns of actual memory consumption per process, then pre-allocate the correct amount of memory and save the user from higher CPU usage.

OPERATING INTEL GRANULATE

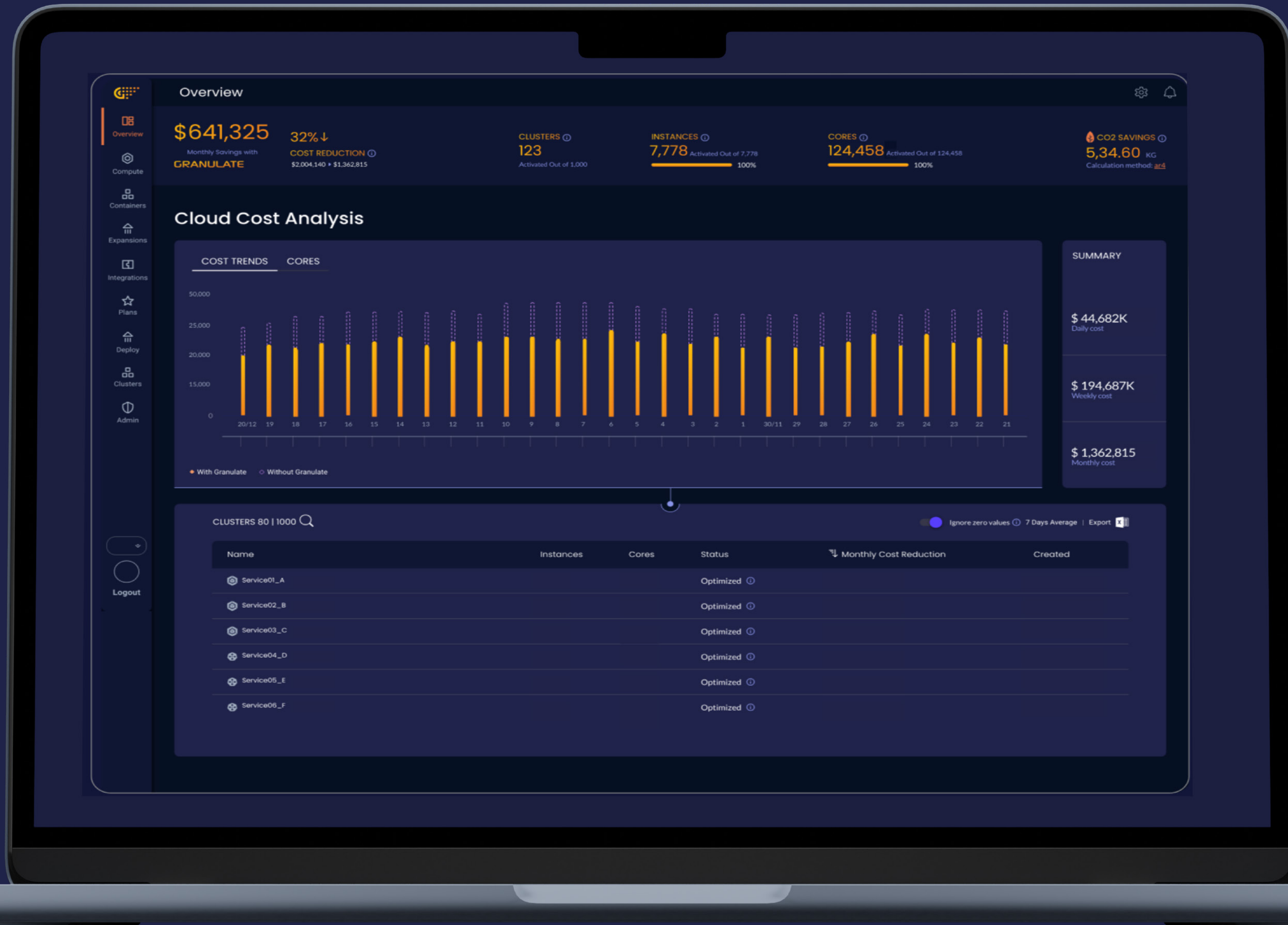
Intel Granulate can be used by technical staff who are responsible for provisioning and infrastructure decisions, and do not need expertise in kernel engineering. A kernel engineer could spend months attempting to optimize just a single configuration, whereas Intel Granulate operates autonomously across a range of configurations, reducing CPU usage without needing the expertise of a kernel engineer. There are four stages to working with Intel Granulate:

- Profile the application to understand where overprovisioning is occurring.
- Learn the behavior of the application with actual workload data patterns.
- Optimize CPU usage by lowering required cores, and lower latency with OS and runtime-level resource allocation.
- Enable capacity reduction by improving the efficiency of active resources and removing idle capacity. This leads to reduced costs.

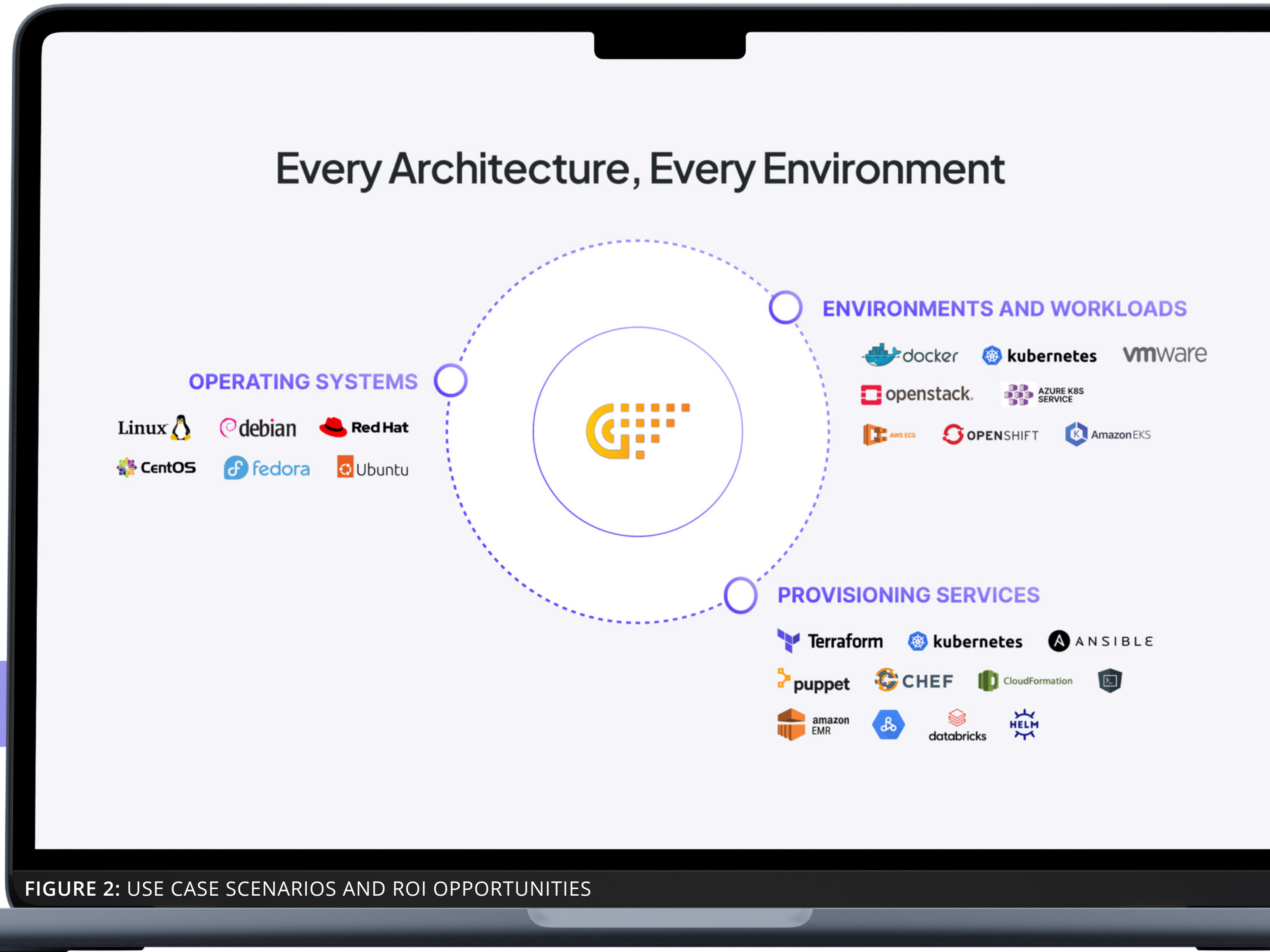
An example of cloud cost analysis is shown in Figure 1 which shows the cost trend over time, with a summary of daily, weekly and monthly costs. Intel Granulate performs capacity optimization followed by rightsizing the allocation to match usage with a recommended buffer allocation (note all thresholds: buffer, max, and min are customizable). All these optimizations are performed automatically. Service providers who must meet SLAs for their customers find it impossible to perform this level of optimization manually and so intentionally overprovision to ensure guarantees are met, thereby increasing their costs.



FIGURE 1:
INTEL GRANULATE
CLOUD COST ANALYSIS



Intel Granulate's ROI opportunities are shown in Figure 2: Different applications yield different optimization opportunities. In Intel Granulate's experience the best opportunities for performance optimization are proprietary Python, Java, Scala, Clojure, Kotlin, Go and Databricks applications, running in containers or as monoliths on cloud instances. The time it takes to learn application behavior also varies with technology, but takes between one to two weeks on average. Intel Granulate only operates on Linux systems, as it connects deeply with the system kernel, and this is only possible with an open-source system.



A FREE PROFILING TOOL FOR THE COMMUNITY

Intel Granulate offers the community an open-source continuous profiling tool, Intel Granulate Continuous Profiler, for analyzing and estimating optimization opportunities. The profiler provides a visual representation of their code in a CPU process flame graph: each function running in the CPU is shown and over time gives the total CPU runtime on the machine. Hovering over a function gives total time %, own time %, and the ability to view specific nodes and containers. Intel Granulate uses the tool to show prospective customers the potential for optimization and cost savings.



Business Benefits

Intel Granulate provides a comprehensive and agile solution to complex challenges resulting in cost-effectiveness, efficiency, and optimized performance at scale.

- **Enhanced Autonomy, Adaptability and Efficiency:** Intel Granulate's dynamic and autonomous resource allocation adjusts the recommended 20% buffer (a customizable threshold) as resource consumption patterns change. This adaptability guarantees smooth performance and responsiveness without manual intervention. Provisioning staff no longer need to err on the side of caution, as Intel Granulate provides transparency into resource usage. This ensures optimal resource allocation without compromising performance.
- **Cost Savings:** By accurately measuring actual resource consumption and dynamically setting a buffer allocation, Intel Granulate helps organizations avoid unnecessary overprovisioning, significantly reducing infrastructure expenses. Intel Granulate helps organizations save on unnecessary computing resources, leading to significant cost reductions.
- **SLA Compliance:** Managed service providers can meet strict SLA guarantees more effectively by avoiding overprovisioning while still having a safety net with the buffer to handle peak demands and contingencies.
- **Transparency and Insight:** Intel Granulate offers valuable insights into resource consumption patterns, empowering organizations to make data-driven decisions for better resource optimization.
- **Scalable Solutions:** Especially for large-scale applications, Intel Granulate's approach can result in substantial cost savings, making it an attractive solution for enterprises with extensive infrastructure needs.
- **Efficient Cloud Native Computing:** Developers can confidently run microservices in containers and leverage Kubernetes without worrying about manual provisioning, allowing them to focus on their core tasks.
- **Real-time Monitoring:** Intel Granulate's real-time monitoring and logging capabilities empower data engineering teams to identify and resolve bottlenecks and fine-tune configurations swiftly, reducing data processing delays.

Call to Action

1 Check your resource's idle time

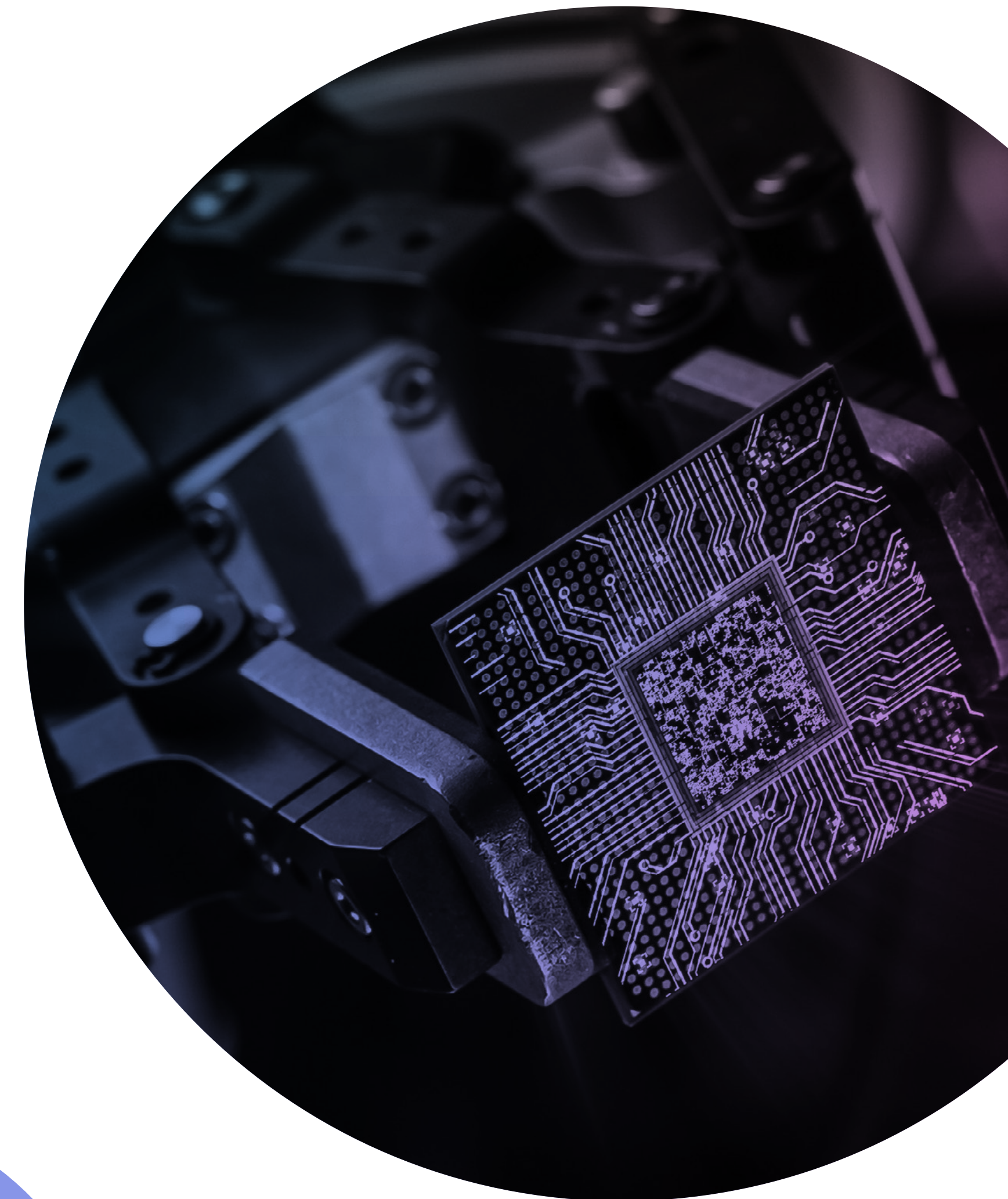
For internet scale operations, where manual scaling is impossible, Intel Granulate offers a workable solution. Furthermore, Intel Granulate can be implemented by technical engineers who are not kernel experts, avoiding hiring requirements to operate Intel Granulate. Omdia suggests your internal FinOps experts report on idle usage, which should indicate a recommended 20% buffer. If it is greater, you should try Intel Granulate to improve your efficiency.

2 Profile your applications

Each use case examined here results in reduced wastage of compute and infrastructure resources. This translates to significant cost savings. Organizations should assess their match to the Intel Granulate use cases, starting with the Continuous Profiler tool.

3 Intel Granulate POC:

Using Intel Granulate frees infrastructure engineers from performing laborious manual optimization tasks, time they could better spend on high-value work. Intel Granulate performs a better job in optimization than any manual method, doing so autonomously and continuously. Omdia recommends organizations engage with Intel Granulate in a proof-of-concept exercise to identify potential cost savings.



Appendix

About



Intel (Nasdaq: INTC) is an industry leader, creating world-changing technology that enables global progress and enriches lives. Inspired by Moore's Law, we continuously work to advance the design and manufacturing of semiconductors to help address our customers' greatest challenges. By embedding intelligence in the cloud, network, edge and every kind of computing device, we unleash the potential of data to transform business and society for the better. To learn more about Intel's innovations, go to newsroom.intel.com and intel.com.

© Intel Corporation. Intel, the Intel logo and other Intel marks are trademarks of Intel Corporation or its subsidiaries. Other names and brands may be claimed as the property of others.

W [Intel.com](https://www.intel.com)

 [Intel](https://twitter.com/intel)

 [Intel](https://www.linkedin.com/company/intel)



Omdia is a global technology research powerhouse, established following the merger of the research division of Informa Tech (Ovum, Heavy Reading, and Tractica) and the acquired IHS Markit technology research portfolio*.

We combine the expertise of more than 400 analysts across the entire technology spectrum, covering 150 markets. We publish over 3,000 research reports annually, reaching more than 14,000 subscribers, and cover thousands of technology, media, and telecommunications companies.

Our exhaustive intelligence and deep technology expertise enable us to uncover actionable insights that help our customers connect the dots in today's constantly evolving technology environment and empower them to improve their businesses – today and tomorrow.

*The majority of IHS Markit technology research products and solutions were acquired by Informa in August 2019 and are now part of Omdia.



Intel Granulate empowers enterprises and digital native businesses with real-time, continuous application-level performance optimization and capacity management, on any type of workload, resulting in cloud and on-prem compute cost reduction. Available in the AWS, GCP, Microsoft Azure and Red Hat marketplaces, the AI-driven technology operates on the runtime level to optimize workloads and capacity management automatically and continuously without the need for code alterations.

Intel Granulate offers a suite of optimization solutions, supporting containerized architecture, big data infrastructures, such as Spark, MapReduce, and Kafka, as well as resource management tools like Kubernetes and YARN. Intel Granulate provides DevOps teams with optimization solutions for all major programming languages, such as Python, Java, Scala, and Go. Customers are seeing improvements in their job completion time, throughput, response time, and carbon footprint while realizing up to 45% cost savings.

DISCLAIMER

Intel Granulate technologies may require enabled hardware, software or service activation. No product or component can be absolutely secure. Your costs and results may vary.

W [Granulate.io](https://granulate.io)

 [Granulate](#)

 [Granulate](#)

 [Granulate](#)

 [Granulate](#)

 [Granulate](#)



Author

The Omdia team of 400+ analysts and consultants are located across the globe

Americas

Argentina
Brazil
Canada
United States

Asia-Pacific



Australia
China
India
Japan
Malaysia
Singapore
South Korea
Taiwan

Europe, Middle East, Africa

Denmark
France
Germany
Italy
Kenya
Netherlands
South Africa
Spain
Sweden
United Arab Emirates
United Kingdom

Omdia

E insights@omdia.com
E consulting@omdia.com
W omdia.com

 [OmdiaHQ](#)
 [Omdia](#)

Citation Policy

Request external citation and usage of Omdia research and data via citations@omdia.com

Michael Azoff



Chief Analyst, Cloud
and Data Center
Research Practice
askananalyst@omdia.com

COPYRIGHT NOTICE AND DISCLAIMER

Omdia is a registered trademark of Informa PLC and/or its affiliates. All other company and product names may be trademarks of their respective owners. Informa PLC registered in England & Wales with number 8860726, registered office and head office 5 Howick Place, London, SW1P 1WG, UK. Copyright © 2022 Omdia. All rights reserved. The Omdia research, data and information referenced herein (the "Omdia Materials") are the copyrighted property of Informa Tech and its subsidiaries or affiliates (together "Informa Tech") and represent data, research, opinions or viewpoints published by Informa Tech, and are not representations of fact. The Omdia Materials reflect information and opinions from the original publication date and not from the date of this document. The information and opinions expressed in the Omdia Materials are subject to change without notice and Informa Tech does not have any duty or responsibility to update the Omdia Materials or this publication as a result. Omdia Materials are delivered on an "as-is" and "as-available" basis. No representation or warranty, express or implied, is made as to the fairness, accuracy, completeness or correctness of the information, opinions and conclusions contained in Omdia Materials. To the maximum extent permitted by law, Informa Tech and its affiliates, officers, directors, employees and agents, disclaim any liability (including, without limitation, any liability arising from fault or negligence) as to the accuracy or completeness or use of the Omdia Materials. Informa Tech will not, under any circumstance whatsoever, be liable for any trading, investment, commercial or other decisions based on or made in reliance of the Omdia Materials.

