SOLUTION BRIEF

GOOGLE DISTRIBUTED CLOUD VIRTUAL EXCELS ON SUPERMICRO® SUPERBLADE®

Simplify hybrid and multi-cloud environments with GDC Virtual on SuperBlade

Google Distributed Cloud

Supermicro SuperBlade with AMD EPYC™ processors powers Google Distributed Cloud Virtual deployments on-premises

Table of Contents

Modern Apps Demand Cloud Native Platforms

The transformative impact of the cloud on businesses has prompted a rapid migration and adoption of cloud-first strategies. As a result, many enterprises are looking into application modernization strategies to meet customer expectations, keep business operations agile, and accelerate innovation. Cloud-native application development empowers enterprises to capitalize on the full power of the cloud by delivering faster time to market, improved efficiency, increased scalability, and better consumer experiences while optimizing cost compared to legacy, on-premise development infrastructures. A cloud-native approach accelerates the application development lifecycle by consuming independent components called microservices, which break large monolithic applications into smaller components.

Cloud-native technologies empower organizations to build and run scalable applications in modern, dynamic public, private, and hybrid cloud environments. Modern cloud-native applications evolve rapidly; leveraging a hybrid cloud platform helps integrate variants of applications across different cloud platforms and/or on-prem environments.

Google Distributed Cloud Virtual (GDC Virtual) is a hybrid cloud management platform that manages containers across multiple cloud platforms and legacy VM workloads.

Businesses can leverage the benefits of GDC Virtual by deploying it on a self-managed Supermicro® SuperBlade® environment to directly run cloud-native applications with CPU, GPU, and other hardware resources.

Drivers of On-Premises Cloud

The cloud has recently become popular for deploying new and existing applications. The simplicity and ease of using a service can offer significant time and cost optimization compared to building your own infrastructure. Many organizations are now adopting cloud-first strategies to leverage these advantages.

Cloud-native approaches are also increasing in popularity and encompass a broadening array of use cases from refactoring existing applications and/or creating container-based microservices architectures. The cloud can even run legacy applications while offering cost savings compared to on-prem deployment models.

Expanding cloud use cases empowers cloud-first organizations to run most of their application landscapes in the cloud, with fewer non-migratable exceptions. However, there is no one simple "lift-and-shift strategy" that meets the needs of every app. Thus, no large organization currently relies solely on modern, cloud-based applications; they must work the cloud into and around existing systems and applications, some of which do not support immediate migration to the cloud. These applications are often the mission-critical systems at the heart of an organization's commercial operations.

The key drivers behind deciding to deploy an on-premises cloud may include:

- **Data security, compliance, and data sovereignty requirements:** Data sovereignty, security, and/or similar restrictions in place because of laws, security policies, or compliance rules may prohibit an application from running in the public cloud. This prescription often extends to Personally Identifiable Information (PII), medical, and/or other sensitive data.

- **Monolithic application design:** Some legacy application architectures don't align with cloud computing pricing models, and the resulting costs and timelines of refactoring prevent organizations from migrating to the cloud.

- **Demand for very low networking latency:** Highly transactional. Low-latency application systems (e.g., banking or transportation) take an unacceptable performance hit if they are too far away from their users, data, or the next-hop data processor in the application flow.

- **Protecting legacy infrastructure investments: Enterprises often look to optimize costs by** leveraging their existing data center investments in servers, networking equipment, and storage devices. Migrating from CapEx to OpEx is simply not viable when the application is economically best served on existing on-premises infrastructure.

One application modernization strategy leverages on-premises cloud solutions such as GDC Virtual to extend cloud capabilities and services to an on-premises data center. This allows the application to leverage many of the advantages of cloud computing while maintaining consistent operations across locations.

GDC Virtual is especially interesting to organizations already using Google Cloud Platform while seeking a seamless solution for integrating their remaining on-premises applications into their cloud operations framework. GDC Virtual empowers organizations to derive the benefits of cloud-based architectures while significantly reducing the costs associated with a DIY

approach across operations, lifecycle management, monitoring, and visibility—all traditionally difficult aspects of IT operations.

GDC Virtual can deploy both traditional and cloud-native applications. Figure 1 shows a single GDC Virtual cluster supporting deployments across multiple cloud platforms, such as Amazon AWS, Google Cloud, and Microsoft Azure. GDC Virtual also includes a bare metal deployment option that delivers many cloud benefits to self-managed Supermicro SuperBlade servers.

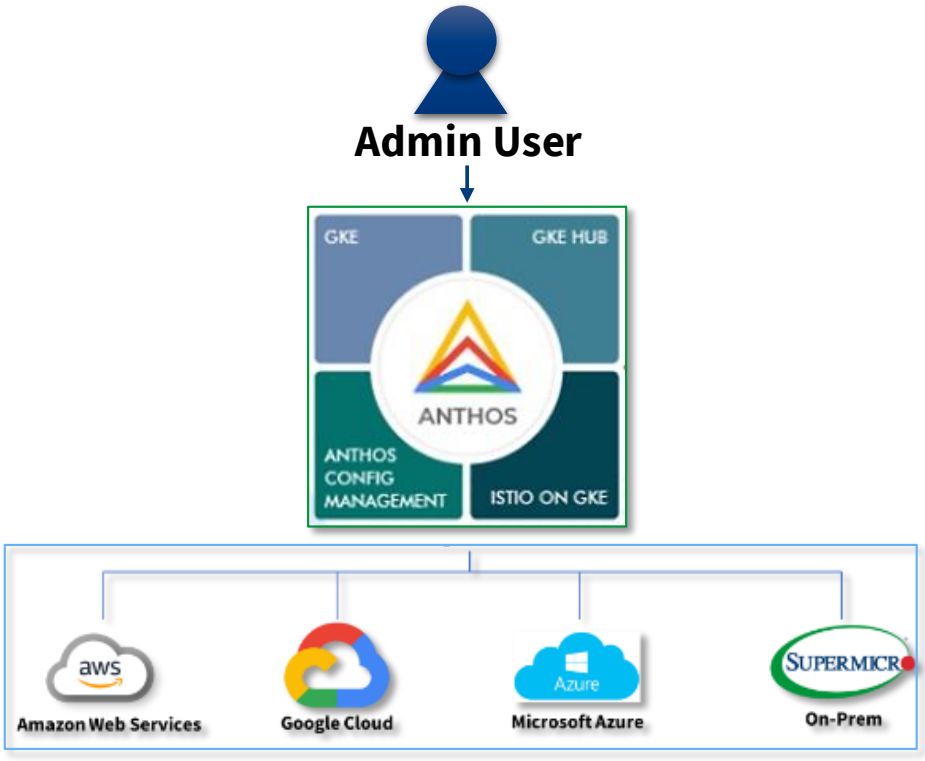

Figure 1 – Sample Google GDC Virtual Deployment.

Solution Overview

Supermicro SuperBlade servers deliver advanced Bare-Metal-as-a-Service (BMaaS) solutions for AI inference, visual computing, Data Center, Cloud, Enterprise, Big Data, and HPC markets. SuperBlade is an ideal platform for running GDC Virtual on bare metal. This Solution Brief presents a solution stack that allows you to deploy a cloud-managed platform using Supermicro SuperBlade servers powered by GDC Virtual.

SUPERMICRO AMD SUPERBLADE

8U SuperBlade Enclosure

Supermicro's high performance, density optimized, and energy-efficient SuperBlade® can significantly reduce initial capital and operational expenses.

SUPERBLADE MODELS

SBA-4114S-C2N/T2N
OCP 3.0 Mezzanine Cards

SBA-4119SG
GPU /PCIe Cards

The Supermicro SuperBlade is powered by AMD EPYC™ Processors

Supermicro's new-generation blade portfolio helps optimize the TCO of key data center components, such as cooling, power efficiency, node density, and networking management. Supermicro SuperBlade, powered by 3rd Gen AMD EPYC™ processors, is built for the most demanding workloads that require high CPU density and the fastest networking available today in a trusted platform that meets enterprise customer demands for on-prem private/hybrid cloud deployments.

The Supermicro SuperBlade comes in an 8U enclosure that accommodates up to 20 hot-pluggable single socket nodes and delivers high performance with AMD EPYC 7003 Series Processors with AMD 3D V-Cache™ technology, DDR4 3200MHz memory, and fast PCIe® Gen4 I/O. Supermicro offers three SuperBlade models powered by AMD EPYC processors: a SAS model, a SATA model, and a GPU-accelerated model, all of which can be mixed in a single 8U enclosure. The 8U SuperBlade can support up to 40 single-width GPUs or 20 double-width GPUs. SuperBlade SAS/SATA models support AIOM for front I/O, which extends the Open Compute Project 3.0 specification to support a wide range of networking options in a small form factor. The 8U SuperBlade also provides customers with advanced networking options, such as 200G HDR InfiniBand and 25G Ethernet switches.

Each SuperBlade enclosure contains at least one Chassis Management Module (CMM), which allows administrators to remotely manage and monitor server blades, power supplies, cooling fans, and networking switches. SuperCloud Composer (SCC) is a composable cloud management platform that provides a unified dashboard for administering software-defined data centers. SCC can orchestrate cloud workloads via the streamlined industry-standard Redfish API. SCC can also monitor and manage a broad portfolio of multi-generation Supermicro servers from a single pane of glass, including SuperBlade.

The AMD EPYC 7003 Series Processors with AMD 3D V-Cache technology are built around the "Zen3" core and contain up to 64 cores per socket and deliver breakthrough per-core performance with 3X the L3 Cache (768 MBs per socket) of general purpose AMD EPYC 7003 CPUs. SuperBlade, powered with AMD EPYC 7003 processors, enables exceptional HPC performance thanks to high frequencies, core counts, high memory bandwidth and capacity, and 768MB of L3 cache.

Supermicro GPU SuperBlade SBA-4119SG supports 3rd Gen AMD 7003 Series EPYC with 3D V-Cache technology processors and the AMD Instinct™ MI210 accelerators. These GPU-accelerated SuperBlade servers are ideal for running AI inference, visual computing, and HPC workloads. SuperBlade systems help organizations reduce time-to-solution for a wide range of applications, add advanced security features, and allow all workloads to run either on-prem or in a public or private cloud. Supermicro SuperBlade offers high density, excellent performance, high power efficiency, and low Total Cost of Ownership (TCO).

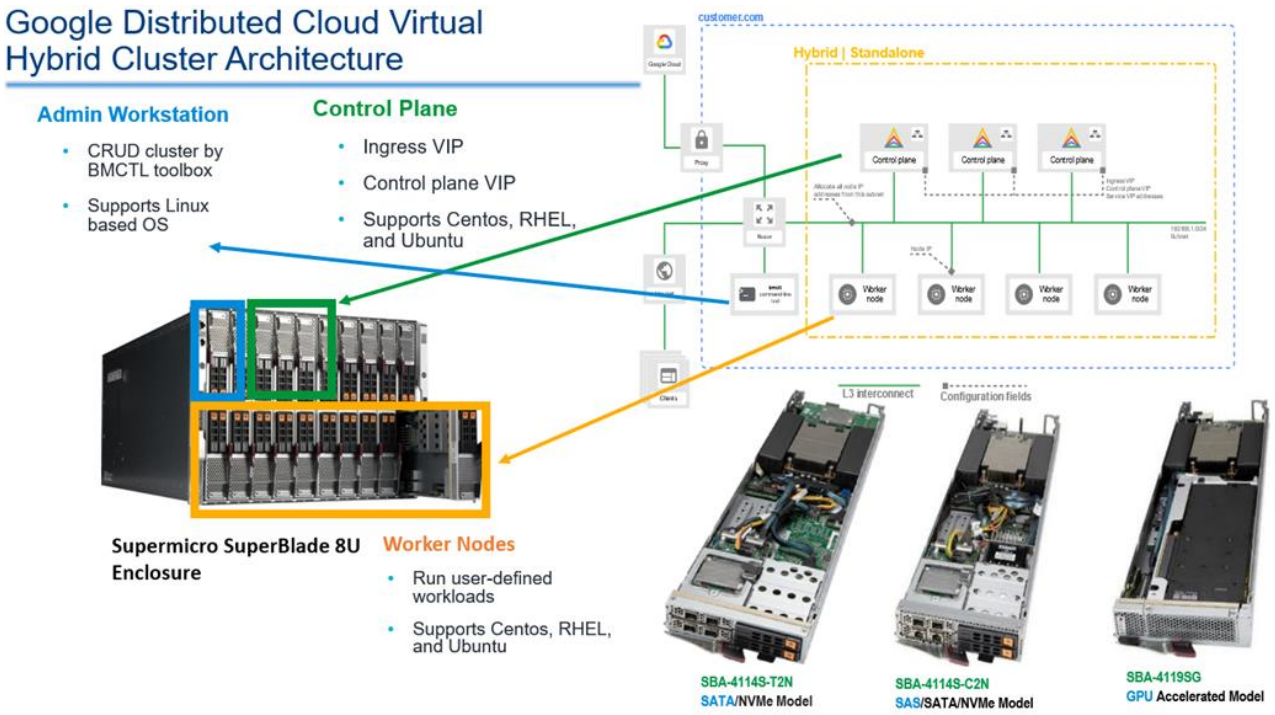

Figure 2 – Google GDC Virtual Hybrid Cluster Architecture deployment on Supermicro SuperBlade.

Figure 2 shows a typical GDC Virtual hybrid cluster architecture deployment on a Supermicro SuperBlade platform where customers can choose SAS, SATA, and GPU accelerated SuperBlade nodes in an 8U enclosure. SuperBlade offers a "private cloud in a box feature" where you can mix and match CPU-only and GPU accelerated nodes in the same 8U enclosure. Customers can use SuperBlade to deploy the admin workstation that hosts command-line interface (CLI) tools and configuration files to provision clusters during installation and CLI tools for interacting with provisioned clusters post-installation. SuperBlade can also host the Control Plane node that includes the Kubernetes API server, etcd storage, and other controllers. You can host the worker nodes that run the actual cloud native applications on CPU- and/or GPU-powered SuperBlade as needed for your workload(s).

Workload Examples

You can quickly deploy various workloads (AI inference, visual computing, 5G/Edge, or any other cloud-native application) in your on-prem environment using Supermicro SuperBlade servers with GDC Virtual via the GKE console. AMD GPU accelerated SuperBlade can be used to perform performance intensive tasks such as AI inference and large scale data processing. GKE offers GPU-specific features, such as time-sharing and multi-instance GPUs, that can improve the efficiency with which your workloads use the GPU resources on your nodes. Local users can visit services via exposed cluster ports. Google Container Registry also provides different images to help rapid deployment and securely manage private images. GDC Virtual on bare metal clusters sends regular health messages to update the health status on Google Kubernetes Engine. Google Cloud Console can display detailed information and perform further operations, such as deploying and deleting workloads.

NGINX is a well-known, free, and open-source web server that can also function as a reverse proxy, load balancer, and HTTP cache. Figure 3 shows a sample NGINX deployment on bare metal SuperBlade servers powered by GDC Virtual.

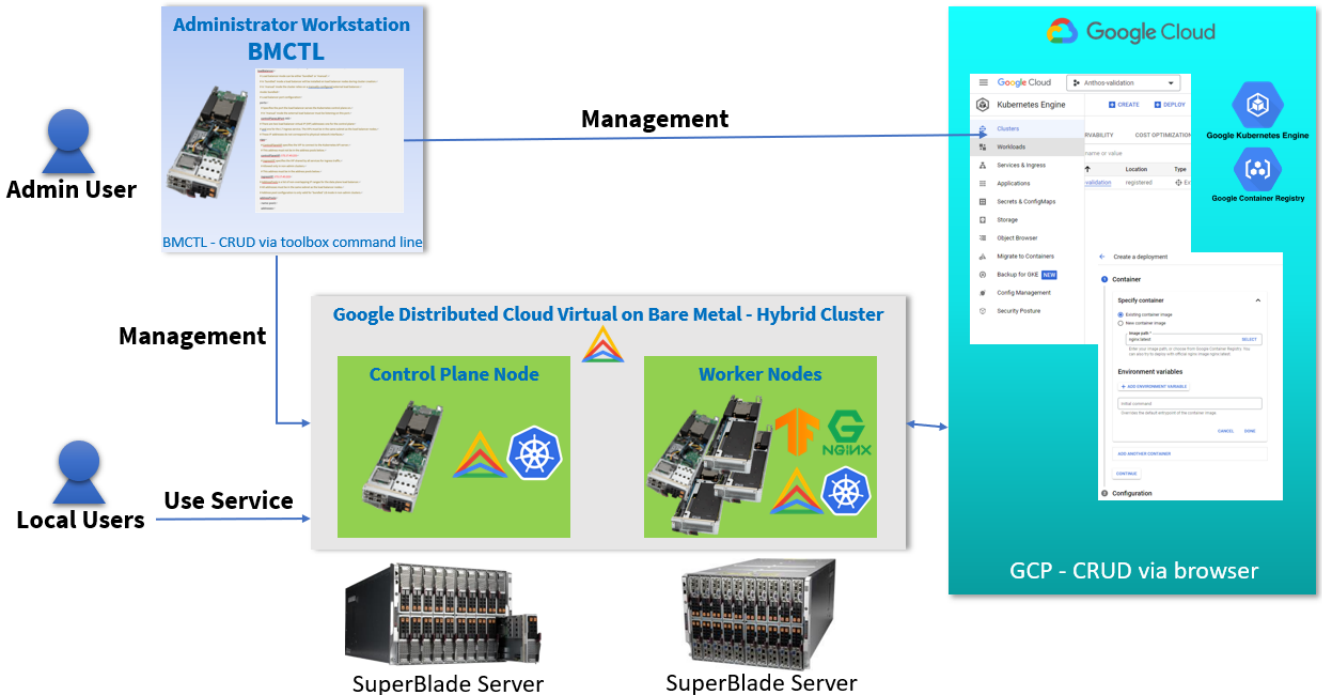

Figure 3 – Sample workload deployment of GDC Virtual on SuperBlade bare metal

Solution Benefits:

Deploying GDC Virtual on Supermicro SuperBlade offers the following benefits:

- Hardware agnostic: Customers can leverage existing on-prem SuperBlade servers to drive data center efficiency.
- No hypervisor layer overhead: Deploying GDC Virtual on SuperBlade has reduced complexity.
- Rapid deployment: Both developers and dev-ops teams benefit from increased productivity because GDC Virtual enables rapid cloud native application development and delivery.
- Superior performance: Supermico SuperBlade, powered by AMD EPYC CPUs and/or AMD Instinct accelerators plus the advanced networking described above, delivers the performance needed to run today's and tomorrow's most demanding workloads on GDC Virtual. A large L3 Cache also reduces memory bandwidth pressure and reduces application latency barriers.
- Easy manageability: SuperBlade CMM manageability coupled with GDC Virtual management enables increased operational efficiency (offers a dashboard with a range of health checks, logging, and monitoring).
- Optimal TCO: Supermicro SuperBlade offers a building block solution with a Resource Saving Architecture that optimizes precious data center resources, such as power, space, and cooling.
- Scalability: SuperBlade enables GDC Virtual to deploy at scale across development, test, and production clusters.

Key Takeaways and Business Benefits

Organizations that need to keep some applications on-premises can turn to Supermicro SuperBlade servers running GDC Virtual on bare metal to transform their on-premises data center into a fully managed, fully integrated cloud region. GDC Virtual is a cloud-native application modernization platform that helps organizations move applications from legacy environments to container-based microservice architectures to reap the full benefits of cloud computing for both cloud-native and legacy applications. GDC Virtual offers a ready-to-go, on-premises platform that helps optimize setup and operating costs while extending the power of Google Kubernetes Engine to Supermicro SuperBlade servers, virtualized on-prem environments, and other public clouds. The Supermicro SuperBlade GDC Virtual solution is ideal for cloud-native workload management. Supermicro SuperBlade servers can be used as control panel nodes and worker nodes to create a GDC Virtual hybrid cluster. GDC Virtual on SuperBlade delivers consistent management, robust security features, out-of-the-box observability, and more. Supermicro SuperBlade, powered by 3rd Gen AMD EPYC processors, runs GDC Virtual with the operational efficiency and consistency needed to meet various SLAs and IT initiatives and empower increased productivity, optimized TCO, and scalability in your data center.

BENEFITS OF DEPLOYING GDC VIRTUAL ON SUPERMICRO SUPERBLADE

Improved productivity for development and security

- Faster application development, testing, and deployment. Developers can write once and deploy anywhere.
- Consistent, unified security policy creation and enforcement.

Streamlined Operational Efficiency

- Reduced complexity with simplified management.
- Faster migrations – IT can quickly containerize, lift and shift applications.
- Reduced effort for releases and patching.

Greener environmental impact and lower TCO

- SuperBlade offers a high-density platform with 96% efficient power supplies there by reducing the carbon footprint.
- Cost optimized building block solution for GDC Virtual with lowest TCO.

As a Google Distributed Cloud Virtual Ready Platform Partner, Supermicro offers a GDC Virtual Ready Platform to run GKE on Bare Metal. We offer a robust application modernization strategy with GDC Virtual on SuperBlade. Deploying GDC Virtual on a tested and validated SuperBlade yields faster time-to-market for new features and increased customer revenue by increasing release frequencies compared to legacy approaches. GDC Virtual on bare metal SuperBlade servers helps future-proof applications by placing them adjacent to Google's cloud services, including advanced machine learning and artificial intelligence options. SuperBlade offers high density, high performance, optimum power efficiency, and ideal Total Cost of Ownership (TCO) for fast, power efficient GDC Virtual deployments on bare metal.

SUPERMICRO

As a global leader in high performance, high efficiency server technology and innovation, we develop and provide end-to-end green computing solutions to the data center, cloud computing, enterprise IT, big data, HPC, and embedded markets. Our Building Block Solutions® approach allows us to provide a broad range of SKUs, and enables us to build and deliver application-optimized solutions based upon your requirements.