BUILDING BIG DATA-AS-A-SERVICE USING CONTAINERS

WHITE PAPER

EXECUTIVE SUMMARY

In a rapidly evolving data center ecosystem that is cluttered with applications, hardware and software, “application-aware” approach is changing the way the data center components interact with each other for better agility, elasticity and scalability. The need to spin up compute and storage capacity independently of each other at the drop of a hat, or the need to scale back capacity during the lean season, or manage the application lifecycle, is a nightmare that most enterprises are experiencing as the current data center architecture has outlived the growth of applications and their capacity needs. With organizations reeling from rising OPEX and CAPEX costs, unmanageable hardware sprawl, and lack of automation for the entire stack, the application-centric approach alleviates the pain on multiple fronts.

THE DATA CENTER EVOLUTION

The traditional hardware-defined data center

Also known as a “siloed” data center, the traditional data center relies heavily on hardware and physical servers. Typically defined by the physical infrastructure dedicated to a singular purpose that determines the amount of data that can be stored and handled by the data center as a whole, the traditional data centers are restricted by the size of the physical space in which the hardware is stored.

Storage, which relies on server space, cannot be expanded beyond the physical limitations of the space. More storage requires more hardware, and filling the same square footage with additional hardware means it will be more difficult to maintain adequate cooling. Therefore, traditional data centers are heavily bound by physical limitations, making expansion a major undertaking.

Software-defined APIs differ across application IO path

Both the infrastructure and application teams manually script workflows specific to each use case, and need separate tools and skill sets for different infrastructure technologies. These gaps place a burden on business agility and drive up costs even more.

Applications are bound to specific resources

Because there is no way to guarantee IOPS in a multi-tenant shared environment, more often than not, utilization is low in anticipation of peak times, noisy neighbor or denial of service. Workflows are app-specific manual scripts.

Application lifecycle management across teams

For example, there is no easy way to clone an entire application. The DB admin needs to know which tables need to be cloned, the application admin then needs to freeze the application state to clone, and the storage admin needs to know which volumes are relevant, as well as free/set cloning space. Nothing is automated, resulting in an uncoordinated mess.

AN APPLICATION-CENTRIC APPROACH

As innovation in the application space continues to accelerate, it is imperative to adopt a technology that treats applications as first class citizens. We need a platform that not only liberates all applications from complex time-consuming deployment processes but also ensures isolation of system resources between all running applications via in-built monitoring. With cloud adoption, the data center has become amorphous, scaling across private and public data centers. Applications now need to be portable and able to move across the boundaries of data centers with the click of a button. This is what Robin Systems delivers. Our goal is to herald the age of the application-driven enterprise.

ASK YOURSELF...

- With your trusted VM technologies - will each hypervisor be aware of neighboring hypervisors and underlying resources?
- Is the hypervisor application-aware and can it support application custom needs?
- Can you easily manage the entire lifecycle for your application/s? This includes tasks such as easy single-click provisioning of distributed clusters, guaranteeing IOPS performance to app-2-spindle RCA, ongoing clones, and snapshots/time-travel.

If you can confidently answer "Yes", you are probably continuously dealing with non-repeatable labor-intensive manual scriptology or, better yet, using a fully automated suite in an optimal manner.

ROBIN CLOUD PLATFORM (RCP)

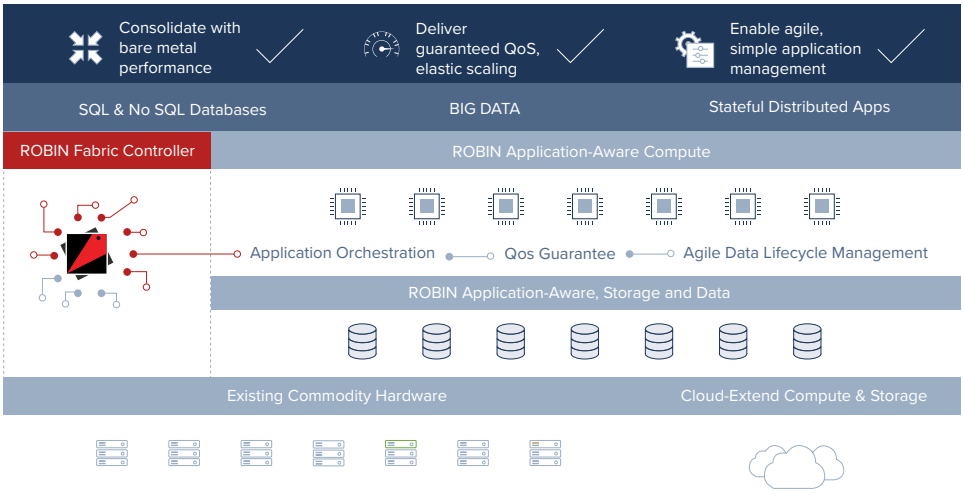
Robin Systems provides a complete out-of-the-box solution for hosting all the applications in an enterprise on a shared platform created out of commodity or cloud components. Robin Cloud Platform can be deployed on bare metal, or on virtual machines, allowing organizations to rapidly deploy multiple instances of their data-driven applications on premise or on cloud, without creating additional copies of data.

RCP Components

Containerized Agile Compute

Container-Aware Scale-Out Storage

Application-Aware Fabric Controller



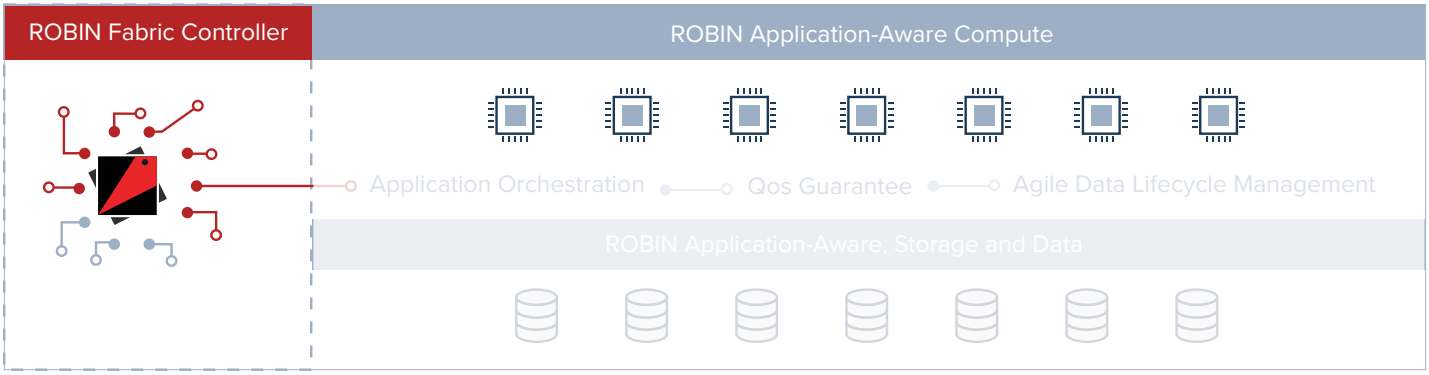
Containerized Agile Compute

RCP leverages container technology to consolidate applications with complete runtime isolation. A container comprises a lightweight, operating system (OS)-based virtualization technology that allows the creation of a compartmentalized and isolated application environment on top of the standard OS. Just as a hypervisor abstracts OS from the underlying hardware, containers help abstract applications from OS and everything underneath, leading to simplified application deployment and seamless portability across hardware and operating platforms.

RCP is the first and only product in the industry that brings containerization benefits to all types of enterprise applications – including highly performance-sensitive workloads such as databases and Big Data.

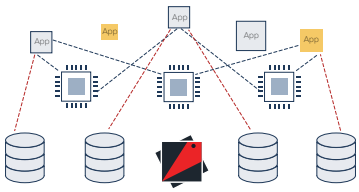
RCP's Adaptive Container technology picks the appropriate container configuration depending on the application types. Traditional applications are deployed within "system containers" to provide VM-like semantics, and RCP supports the deployment of the stateless microservices applications within Docker containers.

When used with bare-metal servers, RCP enables "zero-performance-impact" application consolidation of databases, Hadoop clusters, and other distributed applications such as Elasticsearch, resulting in significant operational efficiency gains and cost reduction.



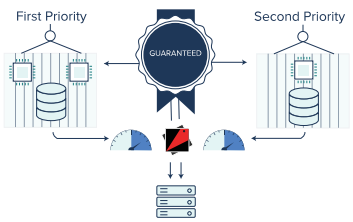
Container-Aware Scale-Out Storage

Traditional storage technology is not designed to handle the rigors of the containerized environment. RCP's container-aware software-defined storage is designed from the ground up to support agile sub-second volume creation, 100,000-plus-variable-sized volumes, and varied file systems and data protection needs of applications running within the containers.



Decoupled Compute and Storage

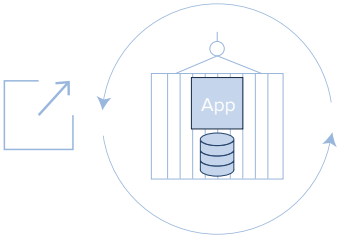
By decoupling storage from compute, RCP ensures data protection against compute failure and enables seamless application mobility. As no persistent data is stored on the compute nodes, the compute layer can be elastically expanded or shrunk without any data movement or copy.



Intelligent Data Protection

RCP "understands" the data protection needs of application components to minimize overhead and maximize performance. For instance, stateful applications such as databases are assigned "protected volumes" using replication whereas the stateless components, such as web servers, are allocated standard volumes. By providing enterprise-grade data protection at the storage layer, RCP obviates the need of inefficient application-level data protection, such as 3-way replication used by Hadoop and other

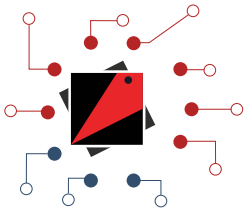
distributed applications. This results in 50 percent or more storage savings and helps improve the write performance.



Application-Driven Data Lifecycle Management

RCP's application-centric data management approach greatly simplifies application lifecycle management. Developers can create quick application clones and snapshots for agile development and testing. Using thin provisioning, compression, and copy-on-write technologies, RCP ensures that clones and snapshots are created in seconds using a minuscule amount of additional storage. Even a 100TB database can be cloned in a snap using only a few GBs of additional storage.

ROBIN Fabric Controller



Application-Aware Fabric Controller

The application-aware fabric controller is the "brain" of the Robin Cloud Platform. This is the component that orchestrates the entire infrastructure fabric from an application perspective. Taking application as the payload, the fabric controller automatically decides the placement, provisions containers and storage for each application component, and configures the application – thus enabling single-click deployment of even the most complex applications.

It also continuously monitors the entire application and infrastructure stack to automatically recover failed nodes and disks, failover applications, and ensures that each application dynamically gets adequate disk IO and network bandwidth to deliver the Application-to-Spindle QoS guarantee.

ROBIN CLOUD PLATFORM USE CASE

Modern applications are built using containers (reflecting application microservices). Each container can be run in a VM or on a bare-metal host. All containers can run on a single host or distributed across many hosts while the storage volumes attached to the containers are directly attached and persistent.

Big Data Use Cases

Provisioning

RCP features such as decoupled compute and storage, data-aware containers, runtime quality of service (QoS) etc. enable single-click provisioning of multi-node Hadoop, Cassandra, MongoDB, Spark, Kafka, etc.

Extensible

Using RCP features such as image and bundle creation workflows, users can define standard application sets such as ELK (Elasticsearch, Logstash and Kibana), or a complicated custom data pipeline, and create end-to-end automation using single-click provisioning.

DevOps

RCP enables users to provision Websphere application servers, Oracle databases, etc. on-the-fly and maintains different versions of a particular product in an image library. Users can stand up or tear down Dev, Test and QA environments instantaneously as required. RCP also supports microservices using Docker for short-lived services. Users can bring their own containers (BYOC) and run it in the RCP environment. Finally, users can containerize and deploy home-grown applications on Robin Cloud Platform.

Multi-tenancy

RCP enables heterogeneous clusters to co-exist and run on the same shared physical hardware. The platform provides runtime management of QoS for clusters with different resource requirements and SLAs.

Hadoop as a Service

RCP enables all components of a data pipeline to be consolidated on a shared platform, and also enables elastic rebalancing of resources across applications. A user can therefore deploy a pipeline composed of Kafka, Hadoop, Spark, Elasticsearch and Cassandra with a single click. Each of these application gets deployed within container-based virtual clusters in just a few minutes.

It is possible to add nodes to a cluster with a few mouse clicks. In addition, RCP's dynamic resource management capabilities allow running clusters to be scaled up to meet the temporary peaks, without requiring data redistribution.

When it comes to Hadoop, RCP provides the following benefits to the different teams including IT infrastructure teams, business users, IT app development teams, management, DevOps, and support teams.

» Agility, flexibility, and elasticity	» Enterprise class high availability
» Single-click cluster provisioning	» Multi-tenancy and data sharing
» Lower cost & reduced hardware footprint	» Very effective monitoring system
» Versatile GUI	» Access to telemetry information

CUSTOMER SUCCESS STORIES

Fortune 500 Retailer CAPEX / OPEX Savings

A real-life use case where this customer in the retail space, has seen the following benefits.

CAPEX Savings

	Before RCP	With Robin Cloud Platform
Servers	Original estimation = 530 traditional servers x $10,000 = $5,300,000 Estimation with RCP = 300 compute and storage servers x $10,000 = $3,000,000	230 fewer servers 230 servers @ $10,000 per server = $2,300,000
Storage	Original estimation 530 servers x 48TB = 25.4PB Estimation with RCP = 2PB = 23.4PB savings	3,846 6TB drives @ $150/drive = $576,900 savings
Software Licences	As per original estimate 530 x $1,000 = $530,000 With RCP = 260 compute nodes x 1,000 = $260,000 = $270,000 savings	Total savings: $3,146,900

OPEX Savings – Hardware and Infrastructure

Operational savings are realized as follows:

- Reduced data center footprint due to 40 percent reduction in number of servers and the associated racks.
- Reduced energy consumption and reduction in energy bill and cooling costs
- Cost savings realized by not having to pay for network port licenses for those 230 machines.
- Much less cabling

OPEX Savings – Software and Operational Efficiency

- Single-click provisioning of Hadoop cluster: Less than 30 minutes instead of the 2-3 weeks required with a traditional non-RCP deployment
- Single resource: One person can handle the deployment of applications and clusters as opposed to 4-5 resources involved in a traditional deployment
- Multi-tenancy: Both Hadoop and non-Hadoop clusters can be provisioned in an instant resulting in a very high degree of agility
- Scalable architecture: Separation of Compute and Storage layer enables independent scaling. This results in cost savings: not paying for extra storage when you need to expand compute capacity only.
- Operational efficiencies with RCP Analytics Module:
- Predict hardware failures such as disk failures well in advance
- Effectively manage QoS resulting in better SLAs and performance
- Get Predictive and What-if analysis using host and container level metrics with descriptive and predictive analytics
- Leverage the highly elastic nature of platform to achieve dynamic expansion of existing application clusters as well as expansion of physical compute and storage layers.

Fortune 500 Retailer

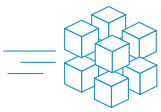
Presented in the table are the execution results of the workload. The job performed better on the Robin Cloud Platform, which included fewer physical nodes and CPU cores.

Feature	Before RCP	With Robin Cloud Platform
Compute	320 cores, 16 nodes	160 cores, 10 nodes
	Dedicated cluster servers	Shared infrastructure
	Manual hardware allocation and fixed infrastructure provisioning	Dynamic container allocation and instantaneous provisioning
Storage	HDD	SSD+HDD
	96 TB/4U	16-24 TB/1u
	Compute tied to Storage	Compute & Storage separation
Execution Time	17 seconds	13 seconds

BENEFITS

Robin Cloud Platform is architected from the ground up to deliver a complete shared platform for hosting all of an enterprise's data and data-driven distributed applications. Some of the key benefits are described below.





Operational Agility & Simplicity	Lower Costs	Better, Predictable Application Performance
Single-click provisioning of clusters and complex distributed applications	CAPEX Reduction – Potential savings of up to 40% with lower HW footprint	Application consolidation with bare metal performance
Push-button cluster extend, application cloning and snapshots	Lower software licensing cost through application consolidation on shared hardware	Automatic Application-to-Spindle performance SLA enforcement

CONCLUSION

Modern data-driven distributed applications need a new kind of infrastructure. An infrastructure that makes machine boundaries functionally invisible. An infrastructure that "understands" the application topology along with the data and is capable of managing data lifecycle. An infrastructure that enables successful multi-tenancy of clustered applications to analyze this data. An infrastructure that gives each application that runs on it complete, dynamic control over its performance and elasticity needs. And an infrastructure that is built with commodity hardware, but provides enterprise-class reliability.