

SCALING CLOUDSTACK TO 100K HOSTS AND MILLIONS OF INSTANCES

Image from freepik.com

Abhishek Kumar

ABOUT ME

- Long-time CloudStack Committer and PMC
- Software engineer @ ShapeBlue
- Father to a delightful 8-month-old baby girl*
- When I'm not working, you'll find me tending

to my ever-growing jungle of houseplants

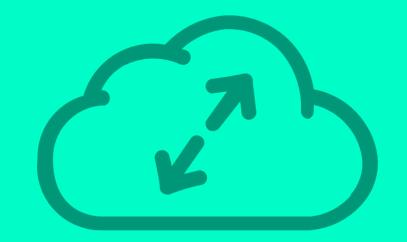


HOW MANY HYPERVISOR HOSTS CLOUDSTACK CAN Support???

How many virtual machines, volumes and other resource?

CLOUDSTACK & Scalability

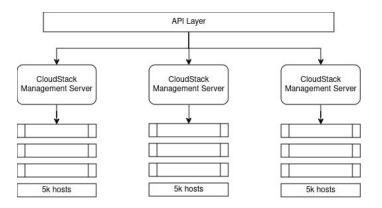
- CloudStack is scalable
- Many users already running production environment with over 5K hypervisor hosts
- Alternative to hyperscalers
 - Identify limits



SCALING THROUGH CELL MODEL

- Multiple CloudStack installations
- Each installation with 5-10k hosts

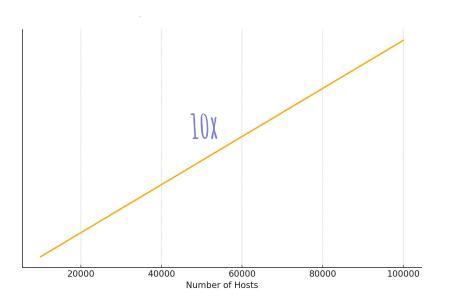
will act as cell



CHALLENGE

Inter-cell communication

SCALE SINGLE INSTALLATION



- Deploy a test environment
 - \circ Simulator based
 - environment
 - Connected agents such as

KVM - How to access

thousands of KVM hosts?

• How to identify issues,

bottlenecks?

- Database
- Concurrency

Management

- Fault Tolerance
- Agent Communication
 - Other issues







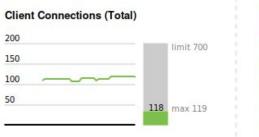
- As the number of resources (e.g., VMs, networks) grow,
 - the database becomes a primary bottleneck
- Centralized database

Use distributed database? A locking service?

DATABASE - CONTD...

MySQL server itself isn't the problem! It can handle upwards of 50k queries per second

- Old structure and usage, old code
- Inefficient usage missing indexes, slow queries, repetitive queries



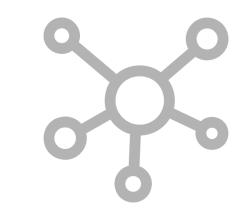
SELECT 9 K/s	
INSERT	CREATE
19 /s	0 /s
UPDATE	ALTER
124 /s	0/s
DELETE	DROP
25 /s	0/s

API HANDLING AND CONCURRENCY MANAGEMENT

- High API request volumes from many users
- Automation scripts and background tasks



- Area where CloudStack has not fared very well in the past
- Intermittent failures shouldn't affect whole environment



AGENT COMMUNICATION OVERHEAD

- Managing many thousands of agents communicating with the CloudStack management server
- Bottlenecks during management server rolling restarts
- Access thousands of hypervisor hosts to create such a large environment



OTHER ISSUES & CONSIDERATIONS

- Resource Allocation and Scheduling
- Periodic maintenance
- Tuning underlying infrastructure, CloudStack features and functionalities
- Security at scale

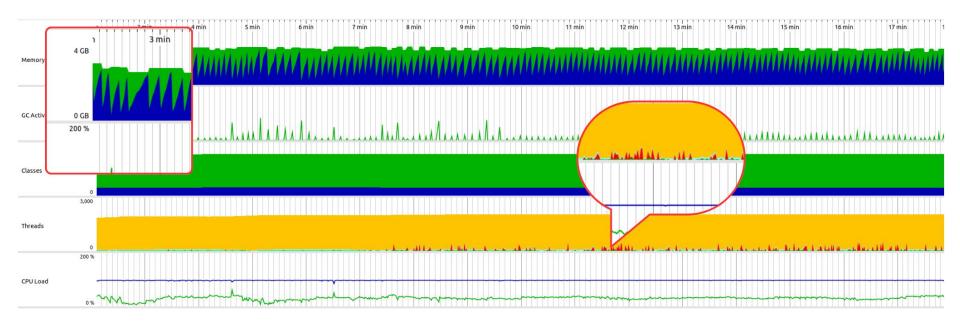


IMPROVEMENTS & CHANGES

- Exhaustive changes but focus on the **persistence layer** and infrastructure resources.
- Primary emphasis on refining **KVM hypervisor** integration.
- Targeted improvements in API and server layers for immediate gains.
- Benchmarking and profiling to identify bottlenecks and achieve peak resource utilization.



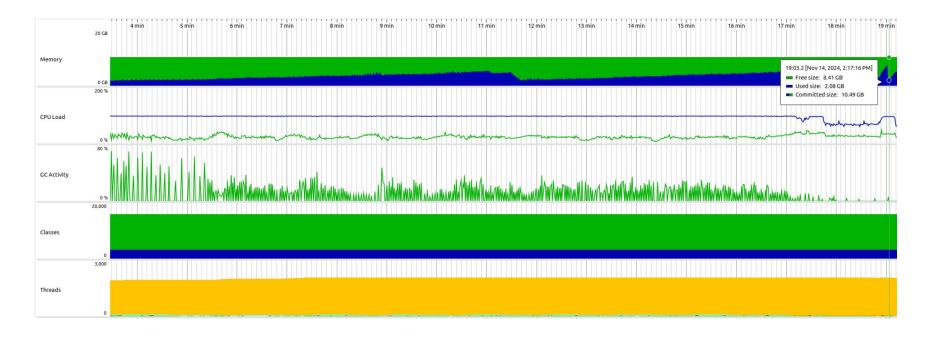
MANAGEMENT SERVER HOST



- Heap size set to 10GB but memory isn't scaling well
- Many threads in blocking state

BEFORE

MANAGEMENT SERVER HOST





- Memory is scaled better
- Blocking threads are minimal

HIKARICP & DATABASE CHANGES

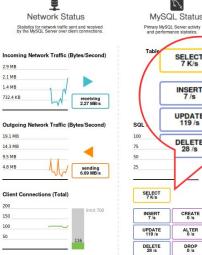
• Default connection pooling library changed from DBCP2 to more performant HikariCP,

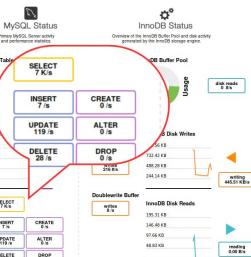
https://github.com/brettwooldridge/HikariCP

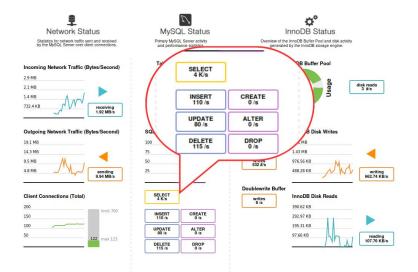
- Option to configure library using db.properties
- Added proper indexes
- Improved JAVA code at different places

Before



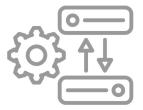






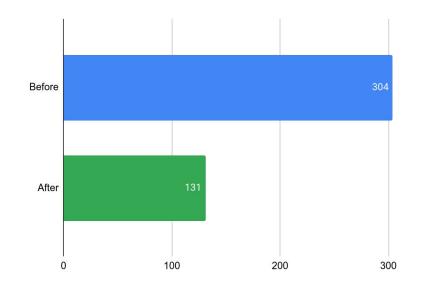
Significant reduction in DB reads

AGENT-SERVER COMMUNICATION



400

- Better handling connections
- Concurrency in TLS/SSL handshakes
- Configuration flexibility
- A **new** mock KVM agent plugin developed



Time taken in milliseconds to reconnect a host

... 2 MOCK KVM AGENT PLUGINS

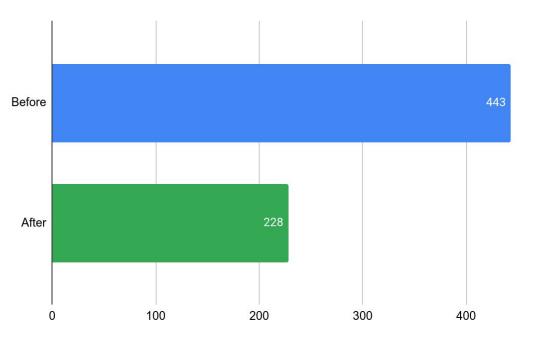
JAVA based using existing agent code

Go-based agent written from scratch

		mai	in.go - cloudstack-go-agent - Visual Studio Code		- ¤ × 🗸	Clusters - 0	EloudStack × +					
File Edi	t Selection View Go Run Terminal Help				*	- > C ()	localhost:5050/#/cluster			00 Q L	e 🖸 🖸	5 I 🗶 E
Ð	EXPLORER	··· 60 CC	ommands.go 👓 🚾 main.go M 🗙 🖘	agent.go <mark>co link.go 1</mark>	ц Ш ··· в	B i 🚨 eConnec	t 📑 CloudStack 🧕 Sofia Jenkins	🕺 Sofia Lab MS 🔞 Staging Rep	o 🧿 👍 [Old UI] Sofia.	🗕 London Jenkins	» I 🗅 A	All Bookmarks
	\sim CLOUDSTACK-GO-AGENT	60 r	nain.go >			20.						
ρ	😅 agent.go		> SetId Aa	ab. 📲 No results 🗠 🗸 🚍 🛪	< L _ /	al al	pachecloudstack	Default vie	w		AU Admir	n User
	👓 answers.go		package main		MEL (9	open source cloud computing					
ĥ	commands.go				Recommendation of the second	🕐 Dashb	oard					
			import (C Baonio	Curu	☆ / Clusters / C Re	fresh	Add cluster +	Search	Q
à	≣ go.sum		" <u>crypto/tls</u> " "crypto/x509"			Compu	ute 🗸	Metrics All	57	Add Gidstor T		~
	ink.go	. 6			Birth Fan	-		All	A			
₿			" <u>io</u> "		All Discourses	Storag	₽ ¥					
	🚥 requesthandler.go		"net" "os"		State Galillani os	E otorug						
Д	👓 responsehandler.go				All control of the second	🔶 Netwo	-t V	Name : Allocation	Cluster type	Hypervisor 2 Hosts	Pod	Zone
			Called Allactor - Monator		Contraction of the second	¬ INELWOI	K T	state	oldster type		name	20116
		13 14				ka Images	5 ×	C2- Enabled	CloudManaged	KVM	POD0	Sanc
		14				Events		KVM	olouumunugou	1 CT III	1000	ount
						Events						
						Project	IS	C3- Enabled	CloudManaged	KVM	PODO	Sanc
								KVM	e le a a la a a a a a a a a a a a a a a			
		20				Roles						
						6		C4- : Enabled	CloudManaged	KVM	PODO	Sanc
		22		= " <u>127.0.0.1</u> " = 8250		Accourt	าเร	KVM	3			
				= 8250 = 0x10000		🔁 Domai	ns					
				= 16384				C5- : O Enabled	CloudManaged	KVM	POD0	Sand
				= " <u>192.168.32.1/24</u> "		f Infrast	ructure ^	KVM	0			
				= 10 = 100								
						EE Su	mmary	C6- Enabled	CloudManaged	KVM	POD0	Sanc
				= 1				KVM				
				= "1:3,1:4,1:5,1:6,1:7"		De Zo	nes					
		32 33		= "/tmp/cloudstack-go-a	igen	E Po	327	4 1			_	,
		34				00 00	as	Showing 1-5 of 5 items	< 1 > 20/pa	ge 🗸		
						品 Clu	usters					
		36 37										
				jger		. Но	sts					
						📄 Pri	mary storage					
8		41 42		zerolog, Infolevel)		IN Se	condary storage					
~~~	> TIMELINE			File(LOG_FILE_PATH, os.0_	CRE		time, otorage	Lice	nsed under the Apa	che License, Version 2.0		
533	> GO		if err != nil {			₽ Sy	stem VMs		CloudStack 4.18.1	.2 O Report issue		<b>\$</b>
× 1	🦻 main* 😌 🛞 0 🛆 3 👷 0	40	Q Ln 6, Col 10	Tab Size: 4 UTF-8 LF ( Go	1.22.0 8							÷

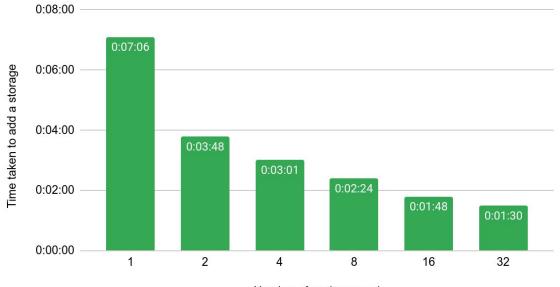
## STORAGE CONNECTIONS

Added concurrency and allow setting worker counts using new global config storage.pool.host.connec t.workers 50k hypervisor hosts present in the zone and set worker count to 2



Time taken in seconds to add a storage

500



Time taken with different worker count comparison

Number of workers used

More workers will not always result in more performance

### CACHING FRAMEWORK

- Not everything can be made faster on its own-sometimes we need caching to bridge the gap
- Caffeine in-memory caching library, <u>https://github.com/ben-manes/caffeine</u>
- Caching added for repeated retrievals:
  - Dynamic Config Keys
  - Account/User role API access



All states 👻 🎇 All thread groups			<ul> <li>Methods</li> </ul>	• •
Hot Spot	Time 👻	Average Time	Events	
SELECT configuration.inistance, configuration.component, configuration.n.         ●       94.2% +42,680 ms +19,533 hot spot ext, oragapatch.commons.dbcp2.D         ▼       94.2% +42,680 ms +19,533 hot spot ext, oragapatch.cloud.util.db.Generic         ▼       94.2% +42,680 ms +19,533 hot spot ext, oragapatch.cloud.util.db.Generic         ▼       94.2% +42,680 ms +19,533 hot spot ext, oragapatch.cloud.util.db.Generic         ▼       94.2% +42,680 ms +19,533 hot spot ext, oragapatch.cloudstack.fra         ▼       94.2% +42,680 ms +19,533 hot spot ext, oragapatch.cloudstack.fra         >       94.2% +42,680 ms +10,531 hot spot ext, org.apatch.cloudstack.acl.Pr         >       94.2% +42,680 ms +10,501 hot spot ext, org.apatch.cloudstack.acl.Pr         >       94.2% +42,680 ms +10,501 hot spot ext, org.apatch.cloudstack.acl.Pr         >       94.2% +42,680 ms +10,501 hot spot ext, org.apatch.cloudstack.acl.Pr         >       94.2% +42,680 ms +10,501 hot spot ext, org.apatch.cloudstack.acl.Pr         >       0.3% +1,481 ms + 10 hot spot ext, org.apatch.cloudstack.acl.Pr         >       0.0% +1,271 µs + 1 hot spot ext, org.apatch.cloudstack.secondar         >       0.0% +1,261 µs + 1 hot spot ext, org.apatch.cloudstack.secondar         >       0.0% +1,261 µs + 1 hot spot ext, org.apatch.cloudstack.secondar         >       0.0% +1,267 µs + 2 hot spot ext. org.apatch.cloudstack.secondar         >       0.	aofase: findByld Chaofase. Icok Row ericDaoBase. FindByld ransactionContextInterceptor.invoke ack. framework.config.ConfigKey.value nework.config.ConfigKey.value jetRioleBasedAPIAccessChecker.checkAccess annicRoleBasedAPIAccessChecker.checkAccess sonaricRoleBasedAPIAccessChecker.checkAccess ystorage.SecondaryStorageManagerimpl.IsPo ystorage.SecondaryStorageManagerimpl.IsPo ystorage.SecondaryStorageManagerimpl.IsPo paredSitatement.excuteQuery chincludingRemoved archincludingRemoved archincludingRemoved JindOneincludingRemoved JindOneincludingRemoved JindOneincludingRemoved Storage.PremotifyParams paoimpl.getConfigValue ystorage.Secondrgvans	ss rimpLscanPool olReadyForScan anStart 2,674 µ		987

Hot Spot SELECT configuration.instance, configuration.component, configuration.n	1.4.4.0.00			•	-	
ELECT configuration.instance, configuration.component, configuration.n	Time	Ψ.	Average Time	E	events	
99.7% - 7.914 ms - 5,340 hot spot evt. com. zaxxer.hikari.pool.HikariProxyPreg     99.7% - 7.914 ms - 5,340 hot spot evt. com. cloud.utils.db. GenericDaoBase.s     9    97.8% - 7.914 ms - 5,340 hot spot evt. com. cloud.utils.db. CenericDaoBase.s     9    97.8% - 7.914 ms - 5,340 hot spot evt. com. cloud.utils.db. CenericDaoBase.s     9    97.8% - 7.914 ms - 5,340 hot spot evt. com. cloud.utils.db. CenericDaoBase.s     9    97.8% - 7.914 ms - 5,340 hot spot evt. com. cloud.utils.db. CenericDaoBase.s     9    97.8% - 7.914 ms - 5,340 hot spot evt. com. cloud.utils.db. CenericDaoBase.s     9    97.8% - 7.914 ms - 5,340 hot spot evt. com. cloud.utils.db. CenericDaoBase.s     9    97.8% - 7.914 ms - 5,340 hot spot evt. com. cloud.utils.db. CenericDaoBase.s     9    97.8% - 7.914 ms - 5,340 hot spot evt. com. cloud.utils.db. CenericDaoBase.s     9    97.8% - 7.914 ms - 5,304 hot spot evt. com. cloud.utils.db. CenericDaoBase.s     9    97.8% - 7.914 ms - 5,304 hot spot evt. com. cloud.utils.db. CenericDaoBase.s     9    97.8% - 7.914 ms - 5,004 hot spot evt. com.cloud.utils.db. CenericDaoBase.s     9    97.8% - 7.914 ms - 500 hot spot evt. com.cloud.event.ActionFventUtils.put     9    1.74% - 7.926 us - 50 hot spot evt. com.cloud.deploy.FirstFitPlaner.rol     9    1.2% - 97.826 us - 50 hot spot evt. com.cloud.dt.eploy.PirstFitPlaner.ls     9    0    1.2% - 97.826 us - 50 hot spot evt. com.cloud.dt.eploy.PirstFitPlaner.ls     9    0    1.2% - 97.435 us - 50 hot spot evt. com.cloud.dt.eploy.PirstFitPlaner.ls     9    0    1.3% - 4.348 us - 50 hot spot evt. com.cloud.dt.eploy.PirstFitPlaner.ls     9    0    0.4% - 4.348 us - 50 hot spot evt. com.cloud.dt.eploy.PirstFitPlaner.ls     9    0.0% - 638 us - 1 hot spot evt. com.cloud.dt.eploy.PirstFitPlaner.ls     9    0.0% - 638 us - 1 hot spot evt. com.cloud.dt.eploy.40a.DomainJoinDabe     9    0.0% - 638 us - 1 hot spot evt. com.cloud.dt.eploy.40a.DomainJoinDabe     9    0.0% - 638 us - 1 hot spot evt. com.cloud.dt.eploy.40a.DomainJoinDabe     9    0.0% - 638 us - 1 h	paredStatement.executeQue searchincludingRemoved Base FindOneIncludingRemoved Base FindOneIncludingRemoved Base FindOneIncludingRemoved Base buildConfigbarams Daoimpl.getConfigValue DishonEventBase Indhangerimpl.getConfigValue DishonEventBaserviceOff derClusters ublishUsageEvent IngManagerimpl.getManagemen ge.PremiumSecondaryStorag aoImpl.searchByTemplateZor JumpLaserchByTemplateZor JumpLaserchByTemplateZor JumpLaserchByTemplateZor JumpLaserchByTemplateZor JumpLaserchByTemplateZor JumpLaserchByTemplateZor JumpLaserchByTemplateZor JumpLaserchByTemplateZor JumpLaserchByTemplateZor	edBy eringNetworkRa ent tState eManagerImpLsc			5,34	0

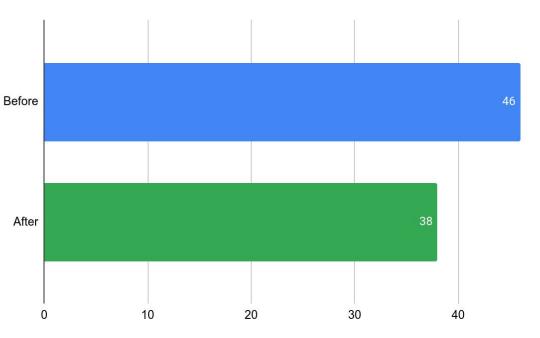


AFTER

- Config value retrievals down to 1/4th
- 30% lesser average retrieval time

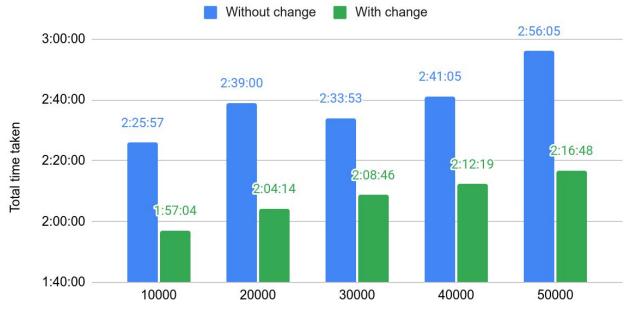
### VM AND DEPLOYMENTS

- While deployment was not the focus but overall changes resulted in gains
- 50k hypervisor hosts
   present in the zone
   and VMs deployed
   with 50 workers



Time taken in seconds to deploy a VM (25k VMs already present)

#### Total time taken to deploy 10k VMs with 50 workers



VMs deployed - 10k added each time

- Good part lesser time after changes
- Bad part linear increase

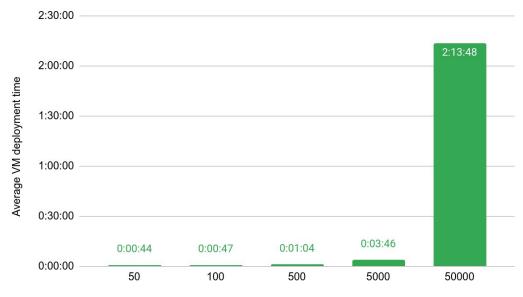




- Identified configurations which can be tuned for optimal performance. These include:
  - Database configurations
  - Global settings
  - $\circ~$  JVM tuning moved to using G1GC instead of ParallelGC
  - OS-level changes#
- Environment recommendations which would allow better performance

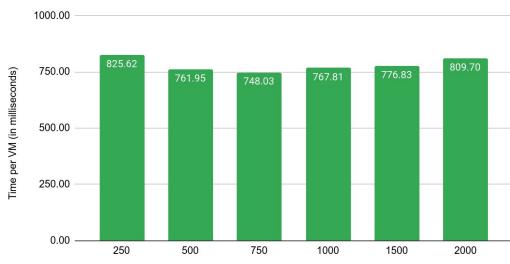
### CLUSTER-HOST BEHAVIOUR

- Performance becomes inversely proportional to number of hosts in a cluster
- # On putting 50k hosts in a single cluster only one VM got deployed



Number of hosts in a single cluster

• Test server was comfortable handling 50-100 workers



Time taken per VM wrt number of VMs deployed

Number of VMs (deployed by workers, 10 VMs by each worker)

### WORK DONE - CONTD...

- 4.20.0
  - Move to more performant HikariCP database connection pooling library, <u>https://github.com/apache/cloudstack/pull/9518</u>
  - Introduced caching framework and dynamic config key caching using Caffeine library, <u>https://github.com/apache/cloudstack/pull/9628</u>
- 4.20.1#
  - Larger scaling work around caching usage, agent-server connection improvements, concurrency, optimisations. <u>https://github.com/apache/cloudstack/pull/9840</u>
  - list*Metrics API related UI improvement,

https://github.com/apache/cloudstack/pull/9825

- Management server maintenance, <u>https://github.com/apache/cloudstack/pull/9854</u>
- $\circ$   $\$  And more...



### FUTURE WORK

- Extend optimization efforts to domains, accounts, VMs, volumes and other resources
- Optimize the allocation and deployment logic
- Develop comprehensive guidelines for architecting and tuning

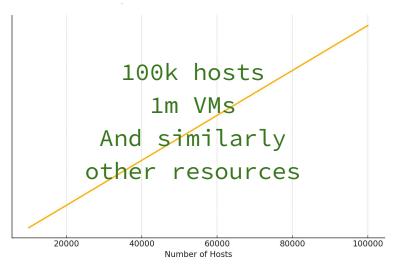
CloudStack environments

- Enhance the efficiency of background tasks
- Incorporating modern design patterns and re-engineering legacy modules

### HOW MANY HYPERVISOR HOSTS CLOUDSTACK CAN SUPPORT???

#### How many virtual machines, volumes and other resource?

apacheclo		E Default view		
② Dashboard		All zones		
Compute	~			
			命 Infrastructure	
Storage	~	Pods	Clusters	
🔶 Network	~	<b>88 1</b>	器 500	
🖾 Images	v	Hosts	Hosts in alert state	
indgoo		50,000	<b>E0</b> 7	
Events				
Projects		Primary storage	System VMs	
		507	Ø 2	
Roles		Virtual routers	Instances	
Accounts		¥ 1,281	@ 113,435	
Domains		.,,		





Thank you!