Openstack Swift at Scale

OPENSTACK SUMMIT 2013 Hong Kong



In the Beginning

- Cloud Files 2.0
- 5 developers
- 4 ops
- 9 months
- 10K lines of code
- Openstack!



Openstack Swift at Rackspace

- 6 Datacenters
- More than 85 Petabytes of raw disk
- Over half a trillion requests since release
- 60 Gb sustained peaks of throughput to a single cluster
 - Over 200 Gb for backend services



The Original Goal

- 100 petabytes of data
- 100 billion objects
- 100 gigabits per second throughput
- 100 thousand requests per second



Use Cases

- Internal and external
- Backup
- Media
- CDN
- Logs



Swift as a Complete System

- Openstack Swift Software
- Hardware
- Network
- Monitoring



Hardware

- Then: 24 1TB drives per box, 1G network
- Now: 90 3TB drives per box, 10G network
- SSD drives for Account/Containers
- Commodity SATA drives
- Test, test, test



Network

- Then: 1GB to host
- Now: 10GB to host
- Network aggregation
- Haproxy for load balancing



Monitoring

- The usual suspects
- Error log lines
- Replication times
- Dispersion report
- Async pendings



Extra Tools

• swift ring master

(https://github.com/pandemicsyn/swift-ring-master)

• swift stalker

(https://github.com/pandemicsyn/stalker)

• graphite/statsd/swift-informant

(https://github.com/pandemicsyn/swift-informant)

• swift spy



💄 Login

Get Nodes

Per zone total writes.

Available graphs: loadavg iops iowait reads writes txbit rxbit gigabyte_used gigabyte_avail service_time util_percentage average_queue_length await

reads for storage1-Z1

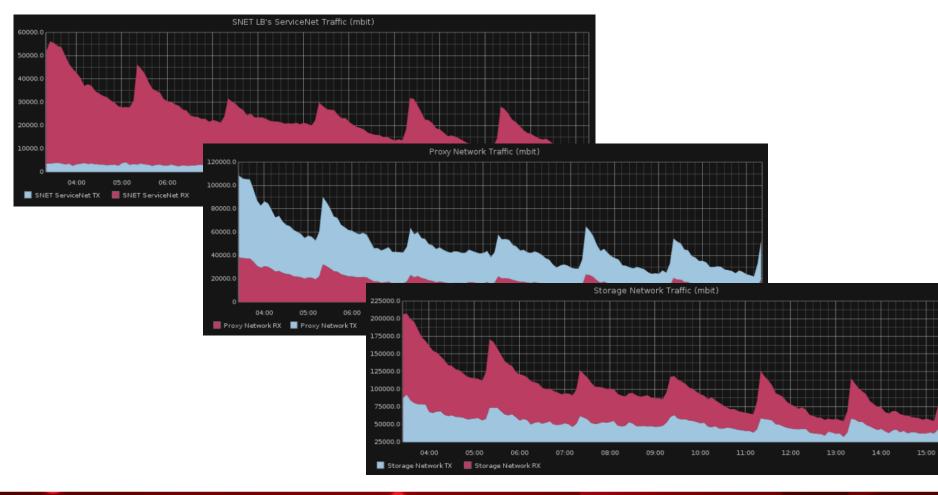
	Oct 27	12 PM	Mon 28	12 PM	Tue 29	12 PM	Wed 30	12 PM	Thu 31	12 PM		
Z1		and the second second		the definition of the	And the second second	white marks and the				7.1k	=	
Z2	a na 17 na mana	्य का जा का सुन् ।	A Pharman the Property of the		readed whether of the "	in the standard by the		Ale references in the second		7.0k		
Z3		া সম্বাদ সম্বাদ্য	and the second start of the second	and the factor of	and Andread and a state of	الدائلة ووقيد				1. h		
Z4		्रिंग स्कृति के स्वयं सहय ह	a later la talende			SwiftSpy				Get Nodes	Login	-
Z5		्राज्य के प्राप्त	a data kata kata kata kata kata kata kat									

Oct 27	12 PM	Mon 28	12 PM	Tue 29	12 PM	Wed 30	12 PM	Thu 31	12 PM
rvers.storage1-Z1 iostat.sda.rea	ads_per_second	hate		last all all		and the second se	and the second		0.012
rvers.storage1-Z1.iostat.sdb.rea	ads_per_second	a da da da anti-	ha minint kan di		ก.ศ.ศักราชการสาวสาวสาวสาว	ndin In Amerika	n staat had a soo nit blood nit	ա. սուսիստ	0.010
rvers.storage1-Z1.iostat.sdc.rea	ads_per_second							i	7.7m
rvers.storage1-Z1.iostat.sdd.rea	ads_per_second	lotal late in the second	الاندار والمتحاد والمتكاف	dos des del tatal desses com	a data menyerin data seri	lines of the shares of the		ւն, , , , , , եել հանդեր	0.099
rvers.storage1-Z1.lostat.sde.rea	ads_per_second	গ্ৰাগ্য নাৰ্							0.011
rvers.storage1-Z1.lostat.sdf.rea	ids_per_second		and the d	a lu la	Induit the L		the control de la	1.101	0.011
rvers.storage1-Z1.lostat.sdg.rea	ads_per_second	real and the second second		And the state of the second second	and the second	According to the	distanti se di di basa da	with the field of the	0.010
rvers.storage1-Z1.iostat.sdh.rea	ads_per_second	and a share to be a share to b				a a teathte i suid a suid a suid	التصفية والمراجع والم	a state and a state of the stat	0.17
rvers.storage1-Z1.iostat.sdi.rea	ds_per_second	land daala nadah fin	Jahran I. (Jahran)	n ha saltailanna stan	ուս են իներել և հերել են	Albert annual del d	ومارياتهم أرواد الشمور والار		0.29
rvers.storage1-Z1.iostat.sdj.rea	dsuper_second	naharanaharah	of the nadel burnling a	mathematics and the		- an third and the short	stands and shift in the	nali bahdarban	0.01
rvers.storage1-Z1.iostat.sdk.rea	ads_per_second	Latin a latera	Later Area Install		م المراجع المراجع الم	منتخبة محمد والقفر	to a factulation and the same		0.01
rvers.storage1-Z1.iostat.sdl.rea	ds_per_second	Ladabhad na ann 16a	eratar, mahil un intera limite			ar maalaha	htelahl matara datah u	նու ստերերի	0.013
rvers.storage1-Z1.iostat.sdm.re	ads_per_second		and the fill						0.01
rvers.storage1-Z1.iostat.sdn.rea	ads_per_second	ded also and the	hilder Handler in	alan a the at han to a	here the second had been been selected as	فليرتجع المتحا المتحا والملا	det la late lla dimensio		0.07
rvers.storage1-Z1.lostat.sdo.rea	ads_per_second	nal de la Californa a	directions controls a sub-	hand die schutzen. In	different al sector de la state de la s	construction dealers are the		and the she will be to be	0.13
rvers.storage1-Z1.lostat.sdp.rea	ads_per_second	فالمرد والمرد والمرد ومرد	and the second second	the second second second second	don ar bitteren an anna a	nala da kara da	and the standard states and the	and the second second	0.05
rvers.storage1-Z1.lostat.sdq.rea	ads_per_second			and the second s		and a second	and an external second	a second second	9.4n
rvers.storage1-Z1.iostat.sdr.rea	ids_per_second	and the first state of the first state of the	distantian data a fi	المراجع المراجع المراجع	di di sentence data di seconda di	والمتعادية المتعادية	a and a large second	dan sa shadaradi sa sa	7.6m
rvers.storage1-Z1.iostat.sds.rea	ads_per_second	and a barrel	CONTRACTOR IN	IL OBLIG	Later a March 19	lite ha ma	a kandulaan u	http://www.com	8.5n
rvers.storage1-Z1.iostat.sdt.rea	ids_per_second	وأعدائها والتلافية	a scied and mail a	dia Intractiti and Albier	والمتعادية والمتعادية والمتعادية	ս առևմել է և և նարդել միս	to lite dial in that a re-	ւր, հրակեսե	0.014
rvers.storage1-Z1.iostat.sdu.rea	ads_per_second	adada Marina ang Kinada	فالقساب فللمع فقع سابقك	et all a factor collection and the second	والمتحافظ والمتحد والمتحاف المحاف	terrest de la companya de la company	landkaran kilikat (- Karolaskaran kilikak ki	and a second of the second	0.05
rvers.storage1-Z1.iostat.sdv.rea	ads_per_second	had a poly by specify and the second s	have distance may a	والمساوية والورجة والمسائب أت	المريار والارتجار المورا الهارك	hals the stand hard line of the	وأرار المريان الماري والمراجع المراجع	hold the Julie of Astron	8.7m
nuare etorana1.71 inetat edw ra	arle nor earond	1					1		0.7



•

OPENSTACK SUMMIT 2013 Hong Kong



OPENSTACK SUMMIT 2013 Hong Kong



The Road Ahead

- Better replication
- Better handling of full disks
- Better error handling/limiting
- Container sync

Thank You!

Chuck Thier Principal Engineer, Rackspace <u>chuck.thier@rackspace.com</u>

@creiht

