Scaling the Boot Barrier: Identifying and Eliminating Contention in OpenStack



Peter Feiner peter@gridcentric.com

Applications as VMs

Applications as VMs

Applications deployed in virtual machines

- Carve up big hosts
- Makes application capacity granular

Applications as VMs

Applications deployed in virtual machines

- Carve up big hosts
- Makes application capacity granular
- Increase capacity by creating more VMs
 - Create more VMs as load approaches capacity
 - When should you create more?

When to Create More

When to Create More

As late as possible

• Avoid over provisioning

When to Create More

As late as possible

- Avoid over provisioning
- As soon as necessary
 - Anticipate when load will surpass capacity
 - Factor in time it takes for new VM start serving
 - How can we optimize this (i.e., make it low)?

- VM Creation Time + Guest preparation time
 - Time for OS to boot and app to start serving
 - Lean OS & stateless app can serve in < 10s
 - Fat OS & big app ready instantly with **live images**

- VM Creation Time +
- Time from nova boot to ACTIVE
- Guest preparation time
 - Time for OS to boot and app to start serving
 - Lean OS & stateless app can serve in < 10s
 - Fat OS & big app ready instantly with live images

- VM Creation Time + Guest preparation time
- Time from nova boot to ACTIVE
- Time for OS to boot and app to start serving

- VM Creation Time + Guest preparation time
- Time from nova boot to ACTIVE
- Time for OS to boot and app to start serving
- Lean OS & stateless app can serve in < 10s

- VM Creation Time +
- Time from nova boot to ACTIVE
- Guest preparation time
 - Time for OS to boot and app to start serving
 - Lean OS & stateless app can serve in < 10s
 - Fat OS & big app ready instantly with live images

- VM Creation Time +
- Time from nova boot to ACTIVE
- Guest preparation time
 - Time for OS to boot and app to start serving
 - Lean OS & stateless app can serve in < 10s
 - Fat OS & big app ready instantly with **live images**



- VM Creation Time +
- Time from nova boot to ACTIVE
- Can take a long time

- Guest preparation time
 - Time for OS to boot and app to start serving
 - Lean OS & stateless app can serve in < 10s
 - Fat OS & big app ready instantly with **live images**



- VM Creation Time +
- Time from nova boot to ACTIVE
- Can take a long time
- Let's do an experiment ...

- Guest preparation time
 - Time for OS to boot and app to start serving
 - Lean OS & stateless app can serve in < 10s
 - Fat OS & big app ready instantly with **live images**



Create VMs in parallel

- Make N creation requests in parallel
- Measure time from API request to ACTIVE

Create VMs in parallel

- Make N creation requests in parallel
- Measure time from API request to ACTIVE
- OpenStack Grizzly
 - Compute: Libvirt + KVM
 - Networking: Quantum + Open vSwitch
 - Storage: qcow2

Create VMs in parallel

- Make N creation requests in parallel
- Measure time from API request to ACTIVE
- OpenStack Grizzly
 - Compute: Libvirt + KVM
 - Networking: Quantum + Open vSwitch
 - Storage: qcow2
- ▶ 96 GB RAM, I2 cores x 2 HT/core, SSD

VM Creation Time

Median Creation Time



VM Creation Time

Median Creation Time



Single VM is fast ~10s



Single VM is fast ~10s



Single VM is fast ~10s

Single VM is fast ~10s



Single VM is fast ~10s

Time (s)

Number Created in Parallel

VM Creation Time

- Single VM is fast ~10s
- Many VMs can be slow
 - Creation time increases linearly with N
 - Must be some bottlenecks

Median Creation Time



Number Created in Parallel

Time (s)

VM Creation Time

- Single VM is fast ~10s
- Many VMs can be slow
 - Creation time increases linearly with N
 - Must be some bottlenecks
- Looks worse without quantum
 - I0s longer when N=20

Median Creation Time



Hardware

• CPUs pegged? RAM all used? Disk busy?

Hardware

• CPUs pegged? RAM all used? Disk busy?

Software

• Locks held for a long time?

Hardware

• CPUs pegged? RAM all used? Disk busy?

Software

- Locks held for a long time?
- Hardware easy to check with atop
 - Let's look at atop first

ATOP -	- node-@	0025904	fe	b5c	201	3/	04/08	15:24:	29						2s	elaps	sed
PRC 1	sys	0.10s		user	0.13s	Ι	#proc	286	5	#:	zomb	ie 0		#exi	t	0	
CPU I	sys	2%		user	5%	Т	irq	09	5 1	i	lle	2401%	5 1	wait		1%	1
CPL I	avg1	0.46	1	avg5	0.16	Т	avg15	0.15	5	C	SW	2058	- I	intr		1103	1
MEM I	tot	62.9G	1	free	58.5G	I	cache	1.80	i I	bı	lff	177.1M	I I	slab	2	5 0.2M	1
SWP I	tot	64.0G	1	free	64.0G	I				V	ncom	1 3.4G	I	vmli	n 🤉	95.4G	1
DSK I		sda	1	busy	1%	I	read	0)	W	rite	e 10) I	avio	3.	20 ms	1
NET I	transpo	ort	1	tcpi	53	I	tcpo	55	5	u	lpi	0) I	udpo		0	1
NET I	network	¢	1	ipi	53	I	ipo	55	5	i	ofrw	ı 0) I	deliv	v	53	1
NET I	lo			pcki	51	I	pcko	51	. I	S		30 Kbps	- I	SO	30	Kbps	1
NET I	eth1			pcki	4	I	pcko	4	- 1	S		1 Kbps	- I	SO	7	Kbps	1
NET I	br100			pcki	4	I	pcko	4	- 1	S	L.	0 Kbps	- I	SO	- 7	Kbps	1
PID	SYSCPL	J USRC	PU	VGROW	RGROW		rddsk	WRDSK	ST	EX	: s	CPUNR	CPU	CMD		1	./2
22089	0.025	s 0.0	4s	0K	0K		0K	16K			- S	22	2%	beam	.sm	ρ	
21367	0.04s	s 0.0	1s	0K	0K		0K	0K			- R	17	2%	atop			
15838	0.015	s 0.0	3s	0K	0K		0K	0K			- S	10	1%	cind	er-v	volume	9
9793	0.00s	s 0.0	2s	0K	0K		0K	0K			- S	1	1%	cind	er-v	volume	9
5180	0.015	s 0.0	0s	0K	0K		0K	20K			- S	1	0%	mysq	ld		
9776	0.015	s 0.0	0s	0K	0K		0K	0K			- S	4	0%	nova	- CO I	nducto	or
9780	0.00s	s 0.0	1s	0K	0K		0K	0K			- S	5	0%	nova	-CO	npute	
9838	0.00s	s 0.0	1s	0K	0K		0K	0K			- S	9	0%	cind	er-v	volume	2
8823	0.00s	s 0.0	1s	0K	0K		0K	0K			- S	9	0%	scre	en		
21552	0.015	s 0.0	0s	0K	0K		0K	0K			- S	4	0%	kworl	ker.	/4:0	

ATOP - node-002590	4feb5c	2013/04/08	15:24:29			2s elapsed			
PRC sys 0.10s	l user	0.13s #pro	c 286 I	#zombie 0	#exi	t 0 I			
CPU I sys 2%	l user	5% lirq	0 %	idle 2401%	l wait	: 1% I			
CPL avg1 0.46	l avg5	0.16 avg1	5 0.15 I	csw 2058	intr	1103 I			
MEM tot 62.9G	free	58.5G cach	e 1.8G	buff 177.1M	slab	260.2M I			
SWP tot 64.0G	l free	64.0G I	I	vmcom 3.4G	vmli	m 95.4G I			
DSK I sda	l busy	1% read	0	write 10	l avio	3.20 ms l			
NET transport	tcpi	53 tcpo	55 I	udpi 0	l udpo) Ø I			
NET network	l ipi	53 ipo	55 I	ipfrw 0	deli	v 53 l			
NET lo	l pcki	51 pcko	51 I	si 30 Kbps	l so	30 Kbps			
NET eth1	l pcki	4 l pcko	9 4 I	si 1 Kbps	l so	7 Kbps			
NET br100	l pcki	4 pcko	9 4 I	si 0 Kbps	l so	7 Kbps			

PID	SYSCPU	USRCPU	VGROW	RGROW	RDDSK	WRDSK	ST	EXC	S	CPUNR	CPU	CMD 1/2
22089	0.02s	0.04s	0K	0K	0K	16K		-	S	22	2%	beam.smp
21367	0.04s	0.01s	0K	0K	0K	0K		-	R	17	2%	atop
15838	0.01s	0.03s	0K	0K	0K	0K		-	S	10	1%	cinder-volume
9793	0.00s	0.02s	0K	0K	0K	0K		-	S	1	1%	cinder-volume
5180	0.01s	0.00s	0K	0K	0K	20K		-	S	1	0%	mysqld
9776	0.01s	0.00s	0K	0K	0K	0K		-	S	4	0%	nova-conductor
9780	0.00s	0.01s	0K	0K	0K	0K		-	S	5	0%	nova-compute
9838	0.00s	0.01s	0K	0K	0K	0K		-	S	9	0%	cinder-volume
8823	0.00s	0.01s	0K	0K	0K	0K		-	S	9	0%	screen
21552	0.01s	0.00s	0K	0K	0K	0K		-	S	4	0%	kworker/4:0

System Wide

	ATOP - node-0025904	feb5c	2013/0	4/08	15:24:29			2s elapsed		
	PRC sys 0.10s	luser (0.13s	#proc	286 I	#zombie 0	#exi	t 0 I		
O	CPU I sys 2%	l user	5% I	irq	0% I	idle 2401%	l wait	1% I		
<u>.</u>	CPL avg1 0.46	l avg5	0.16 I	avg15	0.15 I	csw 2058	l intr	1103 I		
	MEM tot 62.9G	I free 5	58.5G I	cache	1.8G	buff 177.1M	l slab	260.2M		
	SWP tot 64.0G	free 6	64.0G I		1	vmcom 3.4G	l vmli	m 95.4G I		
~	DSK I sda	l busy	1% I	read	0	write 10	l avio	3.20 ms		
	NET transport	l tcpi	53 I	tcpo	55 I	udpi 0	I udpo	0 I		
U U	NET I network	l ipi	53 I	ipo	55 I	ipfrw 0	I deli	v 53 l		
S	NET I lo	l pcki	51 I	pcko	51 I	si 30 Kbps	l so	30 Kbps I		
	NET eth1	l pcki	4 I	pcko	4	si 1 Kbps	l so	7 Kbps I		
S	NET br100	l pcki	4 I	pcko	4	si 0 Kbps	l so	7 Kbps I		
	Π									
	PID SYSCPU USRC	PU VGROW	RGROW R	DDSK	WRDSK ST	EXC S CPUNR CP	PU CMD	1/2		
	22089 0.02s 0.0	4s 0K	0K	0K	16K	- S 22 2	2% beam	.smp		
S	21367 0.04s 0.0	1s ØK	0K	0K	0K	-R 17 2	2% atop			
	15838 0.01s 0.0	3s 0K	ØK	0K	0K	-S 10 1	% cind	er-volume		

0K

0K

0K

0K

0K

0K

0K

ØK

ØК

ØК

0K

ØК

ØК

0K

0K ---

20K --

0K ---

0K ---

0K --

0K --

0K --

- S

- S

- S

- S

- S

- S

- S

1

1

4

5

9

9

4

1% cinder-volume

0% nova-conductor

0% nova-compute

0% cinder-volume

0% kworker/4:0

0% mysqld

0% screen

Per Process

9793

5180

9776

9780

9838

8823

21552

0.00s

0.01s

0.01s

0.00s

0.00s

0.00s

0.01s

0.02s

0.00s

0.00s

0.01s

0.01s

0.01s

0.00s

0K

0K

0K

0K

0K

0K

0K

	CD		025904feb5c			201	04/08	15:24:						2s elapsed					
	CLL.	U	0.10s		user	0.13s	Ι	#proc	286		#zon	bie	0		#exit	5	0	Ι	
<u> </u>	CPU I S	by S	2%	1	user	5%	I	irq	0%	- I	idle		2401%	1	wait		1%	1	
.2	CPL I a	avg1	0.46	1	avg5	0.16	I	avg15	0.15		CSW		2058	- 1	intr		1103	1	
	MEM I t	ot 6	52.9G	1	free	58.5G	I	cache	1.8G	I	buff	1	77.1M	- 1	slab	26	0.2M	1	
	SWP I t	ot 6	5 4.0 G	1	free	64.0G	I			1	VMCC	m	3.4G	- 1	vmlin	n 9	5.4G	1	
2	DSK I		sda	1	busy	1%	I	read	0		writ	e	10	- 1	avio	3.2	0 ms	1	
	NET I t	ranspor	rt	1	tcpi	53	I	tcpo	55		udpi		0	- 1	udpo		0	1	
e E	NETIr	network		1	ipi	53	I	ipo	55		ipfr	w	0	- 1	deliv	/	53	1	
S	NET I 1	.0		1	pcki	51	I	pcko	51	. I	si	30	Kbps	- 1	SO	30	Kbps	1	
	NET I e	eth1		1	pcki	4	I	pcko	4	- I	si	1	Kbps	- 1	SO	7	Kbps	1	
	NET I b	or100		1	pcki	4	I	pcko	4	- I	si	0	Kbps	1	SO	7	Kbps	1	
	PID	SYSCPU	USRC	PU	VGROW	RGROW		rddsk	WRDSK	ST	EXC S	CP	UNR (CPU	CMD		1	./2	
	22089	0.02s	0.0	4s	0K	0K		0K	16K		- 5	5	22	2%	beam.	smp			
S	21367	0.04s	0.0	1s	0K	0K		0K	ØK		- F	2	17	2%	atop				
انه	15838	0.01s	0.0	3s	0K	0K		0K	ØK		- 5	5	10	1%	cinde	er-v	olume	ļ	
Ŭ	9793	0.00s	0.0	2s	ØK	0K		0K	ØK		- 5	5	1	1%	cinde	er-v	olume	ļ	
0	5180	0.01s	0.0	0s	0K	0K		0K	20K		- 5	5	1	0%	mysql	d			
	9776	0.01s	0.0	0s	0K	0K		0K	ØK		- 5	5	4	0%	nova	-con	ducto	r	
	9780	0.00s	0.0	1s	ØK	0K		0K	ØK		- 5	5	5	0%	nova	-com	pute		
	9838	0.00s	0.0	1s	ØK	0K		0K	ØK		- 5	5	9	0%	cinde	er-v	olume	ļ	
<u> </u>	8823	0.00s	0.0	1s	ØK	0K		0K	ØK		- 5	5	9	0%	scree	en			
					(at			р										
-------------	----------------------------------	---	-------------------------------	----------------------	-------------------------------------	-----------------------------	------------------	--------------------------------	----------------------	------------------------	------------------------	--------------------------	----------------------------------	----------------------	----------------------------------	----------------------------	---------------------------------	----	----
de		U sys avg1	025904 0.10s 2% 0.46	fet 	o 5c user user avg5	2013 0.13s 5% 0.16	3/ 	04/08 #proc irq avg15	15: 0.15	6			240170 2058	-	wait	24	40)1	.%
n N N	MEM I SWP I DSK I	tot tot	62.9G 64.0G sda		free free busy	58.5G 64.0G 1%		cache read	1.80		l b I v I w	uff mcom rite	177.1M 3.4G 10		slab vmlin avio	26 1 9 3.2	50.2M 5.4G 20 ms		
Syster	NET I NET I NET I NET I	transpo network lo eth1 br100	rt 		tcpi ipi pcki pcki pcki	53 53 51 4		tcpo ipo pcko pcko	55 51 2) 5 - -	u i s s	apı pfrw i i	0 30 Kbps 1 Kbps 0 Kbps		uapo deliv so so	/ 30 7 7	0 53 Kbps Kbps Kbps		
S	PID 22089 21367	SYSCPU 0.02s	USRO 0.0	PU 94s	VGROW ØK	RGROW ØK		RDDSK ØK	WRDSK 16K	ST	EX	C S - S	CPUNR (22	CPU 2%	CMD beam.	smp	1.)	/2	
oces	15838 9793 5180	0.045 0.01s 0.00s 0.01s	0.0 0.0 0.0)3s)2s)0s	0K 0K 0K	0K 0K 0K		ØK ØK ØK	0K 0K 20K			- S - S - S	10 1 1	2% 1% 1% 0%	cinde cinde mysql	er-v er-v .d	volume volume		
Per Pr	9776 9780 9838 8823	0.01s 0.00s 0.00s 0.00s	0.0 0.0 0.0 0.0	0s 1s 1s 1s	0K 0K 0K 0K	0K 0K 0K		0K 0K 0K 0K	0K 0K 0K 0K	 		- S - S - S - S	4 5 9 9	0% 0% 0% 0%	nova- nova- cinde scree	-con -con er-v en	nducto npute volume	r	
-	21552	0.01s	0.0	0s	ØK	ØK		ØK	ØK			- S	4	0%	kwork	ker/	′4:0		

					at		.0	р										
Vide	CPU MEM	02590 0.10 29 0.40 62.90	04fel s % 6 G	user user avg5 free	201 0.13s 5% 0.16 58.5G	3/ 	04/08 #proc irq avg15 cache	15: 0.15 1.80	5			2401% 2058 177.1M		wait intr slab	24	1103 50.2M)1	%
System V	DSK NET trans NET netwo NET lo NET eth1 NET br100	64.00 sdd sport ork 	6 a - -	tree busy tcpi ipi pcki pcki pcki	64.00 1% 53 53 51 4 4		read tcpo ipo pcko pcko pcko	6 55 51 2	0 5 L 4	Vn wr ud ip si si	ncom rite dpi ofrw i 3 i	3.4G 10 0 0 Kbps 1 Kbps 0 Kbps		vmlin avio udpo deliv so so so	3.2 3.2 30 7 7	20 ms 0 53 Kbps Kbps Kbps		
Per Process	PID SYS 22089 0.0 21367 0.0 15838 0.0 9793 0.0 5180 0.0 9776 0.0 9780 0.0 9838 0.0 8823 0.0 21552 0.0	CPU US 02s 0 04s 0 01s 0 01s 0 01s 0 01s 0 01s 0 01s 0 00s 0 01s 0	RCPU .04s .01s .03s .02s .00s .00s .01s .01s .01s .01s	VGROW ØK ØK ØK ØK ØK ØK	RGROW ØK ØK ØK ØK ØK ØK		RDDSK ØK ØK ØK ØK ØK ØK ØK	WRDSK 16K 0K 0K 0K 20K 0K 0K 0K 0K	ST 	EX(C S C - S - R - S - S - S - S - S - S - S - S - S - S	PUNR 22 17 10 1 1 4 5 9 9 9 4	CPU 2% 2% 1% 1% 0% 0% 0% 0% 0%	CMD beam. atop cinde cinde mysql nova- nova- cinde scree kwork	smp er-v er-v d -cor er-v en ker/	1 /olume /olume nducto npute /olume /4:0	/2 r	

			ć	at		р							
Wide	CPU MEM	025904feb 0.10s 2% 0.46 62.9G 64.0G	Sc user Fr	2013/ 0.13s 5% ee	04/08 #proc	15:i	d 8	le .5	G	 	24 wait intr slab 26 vmlim 9	1103 50.2M)19
System	DSK NET transpo NET network NET lo NET eth1 NET br100	sda ort 	busy tcpi ipi pcki pcki pcki	1% 53 53 51 4 4	read tcpo ipo pcko pcko pcko	0 55 55 51 4 4	wi ua iµ s ⁻ s ⁻	rite dpi pfrw i 30 i 1 i 0	10 0 Kbps Kbps Kbps		avio 3.2 udpo deliv so 30 so 7 so 7	20 ms 0 53 Kbps Kbps Kbps	
Per Process	PID SYSCPL 22089 0.023 21367 0.043 15838 0.013 9793 0.003 5180 0.013 9776 0.013 9780 0.003 9838 0.003 8823 0.003 21552 0.013	USRCPU 0.04s 0.01s 0.03s 0.02s 0.00s 0.00s 0.00s 0.01s 0.01s 0.01s 0.01s 0.01s	VGROW ØK ØK ØK ØK ØK ØK ØK	RGROW ØK ØK ØK ØK ØK ØK ØK	RDDSK ØK ØK ØK ØK ØK ØK ØK	WRDSK S 16K - 0K - 0K - 0K - 20K - 0K - 0K - 0K - 0K - 0K -	T EX(C S CP - S - R - S - S - S - S - S - S - S - S - S - S	UNR (22 17 10 1 1 4 5 9 9 9 4	PU 2% 2% 1% 1% 0% 0% 0% 0%	CMD beam.smp atop cinder-v cinder-v mysqld nova-com nova-com cinder-v screen kworkerv	1 Volume Volume nducto npute Volume Volume	/2 or

					Ċ	at	-	Ο	р)										
Je	CP	U	025904 0.10s 2%	feb5c	r r	2013 0.13s 5%	3/04 #	4/08 #proc	15:	Ļ	d	1	e		!	wait	24	10)1	%
n Wic	ME	M	0.46 62.9G 64.0G sda	†	r	ee	Ir	read		5	2	S , wri	te	10		intr slab vmlin avio	260 195 3,20	1103 0.2M 5.4G 0 ms		
/sten	DS NET 1		ort : 	tcp ipi pck	i	53 53 51	l t l i l p	tcpo ipo pcko		55 55 51		udp ipf si	rw 30	0 0 Kbps		udpo deliv so	/ 30	0 53 Kbps		
Ś		etni br100	 	I pck	i ROW	4 4 RGROW	l p l p RI	DDSK	WRDS	4 4 K S		si si	0 S CP	Kbps Kbps UNR	I CPU	so so	71	Kbps Kbps 1	1	
ess	22089 21367 15838	0.02 0.04 0.01	s 0.0 s 0.0 s 0.0	04s 01s 03s	0K 0K 0K	0К 0К 0К		0К 0К 0К	16 0 0	K - K - K -	-	-	S R S	22 17 10	2% 2% 1%	beam. atop cinde	smp er-ve	olume		
Proc	9793 5180 9776 9780	0.00 0.01 0.01 0.00	s 0.0 s 0.0 s 0.0 s 0.0	02s 00s 00s 01s	0K 0K 0K 0K	0К 0К 0К 0К		0К 0К 0К 0К	0 20 0 0	K - K - K - K -	-	-	S S S S	1 1 4 5	1% 0% 0% 0%	cinde mysql nova- nova-	er-vo ld -cono -com	olume ducto pute	r	
Per	9838 8823 21552	0.00 0.00 0.01	s 0.0 s 0.0 s 0.0	01s 01s 00s	0К 0К 0К	0К 0К 0К		0К 0К 0К	0 0 0	K - K - K -	-	- - -	S S S	9 9 4	0% 0% 0%	cinde scree kwork	er-vo en ker/4	olume 4:0		

					at	\mathcal{O}	р						
de	CP	U	025904f 0.10s 2% 0.46	eb5c user user	2013, 0.13s 5%	/04/08 #proc	¹⁵ i(11	e		24 wait		1%
m Wi	ML DS	:M	62.9G 64.0G sda	hu	ee	-	5	٥.	20 1%		slab 260 vmlim 95 avio 3.20 udpo	.2M .4G ms 0	
Syste	NET I NET I NET I	Lo eth1 br100	; 	PCRL pCRL pCki pCki	4 4	l pcko l pcko l pcko	4 4	si si si	1 Kbps 0 Kbps	1	deliv so 30 K so 7 K so 7 K	53 bps bps bps	
SSS	PID 22089 21367 15838	SYSCP 0.02 0.04 0.01	U USRCP s 0.04 s 0.01 s 0.03	U VGROW s ØK s ØK s ØK	RGROW ØK ØK ØK	RDDSK ØK ØK ØK	WRDSK ST 16K 0K 0K	EXC S - S - R - S	CPUNR 22 17 10	CPU 2% 2% 1%	CMD beam.smp atop cinder-vo	1/2 lume	Γ
r Proce	9793 5180 9776 9780	0.00 0.01 0.01 0.00	s 0.02 s 0.00 s 0.00 s 0.01	s ØK s ØK s ØK	0K 0K 0K 0K	0K 0K 0K 0K	0K 20K 0K 0K	- S - S - S	1 1 4 5	1% 0% 0% 0%	cinder-vo mysqld nova-cond nova-comp	lume luctor oute	
Pe	8823 21552	0.00 0.01	s 0.01 s 0.00	s ØK s ØK	ØK ØK	ØK ØK	0K 0K	- S - S	9 4	0% 0%	screen kworker/4	:0	

Sample every 2s using atop -w log 2

- **Sample every 2s using** atop -w log 2
- HW utilization for N=20:

- **Sample every 2s using** atop -w log 2
- HW utilization for N=20:

Resource	Metric
RAM	% Used
CPU	% Time Busy
Disk	% Time Busy

- **Sample every 2s using** atop -w log 2
- HW utilization for N=20:

Resource	Metric	Median	Max
RAM	% Used	9	11
CPU	% Time Busy	14	55
Disk	% Time Busy	9	80

- **Sample every 2s using** atop -w log 2
- HW utilization for N=20:

Resource	Metric	Median	Max
RAM	% Used	9	
CPU	% Time Busy	14	55
Disk	% Time Busy	9	80

Lots of capacity for parallelism

- **Sample every 2s using** atop -w log 2
- HW utilization for N=20:

Resource	Metric	Median	Max
RAM	% Used	9	
CPU	% Time Busy	14	55
Disk	% Time Busy	9	80

- Lots of capacity for parallelism
 - Time to look at SW

- Anything that inhibits parallelism
 - Some kind of lock contention

- Anything that inhibits parallelism
 - Some kind of lock contention
- Hopefully easy to fix :-)
 - Many locking strategies exist

- Anything that inhibits parallelism
 - Some kind of lock contention
- Hopefully easy to fix :-)
 - Many locking strategies exist
- Identified using tracing
 - Let's take a look

- Record events during application execution
 - e.g., Function entry & exit, lock acquisition

- Record events during application execution
 - e.g., Function entry & exit, lock acquisition
- Visualized as stacked extents:



- Record events during application execution
 - e.g., Function entry & exit, lock acquisition
- Visualized as stacked extents:



Traces are usually pretty busy ...

Tracing OpenStack

Tracing OpenStack

- Added @traced to nova and quantum
 - Events on function call and return
 - Events before and after lock()
 - Outputs to trace-viewer format
 - Using Google Chrome? See <u>about:tracing</u>

Tracing OpenStack

- Added @traced to nova and quantum
 - Events on function call and return
 - Events before and after lock()
 - Outputs to trace-viewer format
 - Using Google Chrome? See <u>about:tracing</u>
- Repeat experiments with tracing on and hunt for bottlenecks
 - Look for stretched extents

Hunting: Resource Accounting

Hunting: Resource Accounting

Resource Accounting

- Enforces max RAM, VCPUs, etc. allocated
- Global lock per compute node

Can add 15s of serialization to VM creation

- Can add 15s of serialization to VM creation
- Slow because of RPC to conductor

- Can add 15s of serialization to VM creation
- Slow because of RPC to conductor
- Solution Part I: Remove NOP updates
 - Reduces median creation time 10% when N=20

- Can add 15s of serialization to VM creation
- Slow because of RPC to conductor
- Solution Part I: Remove NOP updates
 - Reduces median creation time 10% when N=20
- Solution Part 2: Coalesce RPCs
 - Future work

b libvirt starts qemu process, apparmor, etc.

libvirt starts qemu process, apparmor, etc.

Global lock... can't fix this in OpenStack

- b libvirt starts qemu process, apparmor, etc.
- Global lock... can't fix this in OpenStack
- Can we mitigate the problem?

Bottleneck: Libvirt

Bottleneck: Libvirt

- Many short calls (e.g., get hostname)
 - Become long calls due to global lock
Bottleneck: Libvirt

- Many short calls (e.g., get hostname)
 - Become long calls due to global lock
- Solution: avoid unnecessary calls
 - Down from 248 to 7
 - Reduces max creation time 20% when N=20

- eventlet's "green" threads are coroutines multiplexed on single native thread
 - You can't block in a green thread
 - Python's stdlib patched to yield instead of block
 - C libraries aren't patched

- eventlet's "green" threads are coroutines multiplexed on single native thread
 - You can't block in a green thread
 - Python's stdlib patched to yield instead of block
 - C libraries aren't patched
- Pool of native threads to use blocking libs

- eventlet's "green" threads are coroutines multiplexed on single native thread
 - You can't block in a green thread
 - Python's stdlib patched to yield instead of block
 - C libraries aren't patched
- Pool of native threads to use blocking libs
- Maybe there's more room for improvement

• One work queue per worker thread

- One work queue per worker thread
- Green-thread to work-queue map is fixed:

- One work queue per worker thread
- Green-thread to work-queue map is fixed:
- Solution: use a global work queue
 - Get to wait on libvirt lock sooner :'-(







• VM creation time:

- Max 20% lower
- Median 10% lower



• VM creation time:

- Max 20% lower
- Median 10% lower
- Wait for libvirt sooner
 - On the bright side, once libvirt fixed, OpenStack has fewer bottlenecks



Conclusion

• Low VM creation time is good

- Necessary for scaling
- VM Creation time scales poorly due to software contention
 - Bottlenecks in OpenStack code easily fixed
 - libvirt still a big bottleneck
- Tracing helps identify contention

Future Work

- Coalesce RPC updates to conductor
- Eliminate big qemu lock in libvirt
- Instrument other OpenStack services (glance, swift, cinder, etc.)
- Perform more experiments

Questions?



Peter Feiner

peter@gridcentric.com

github.com/peterfeiner/{nova,quantum}/tree/tracing