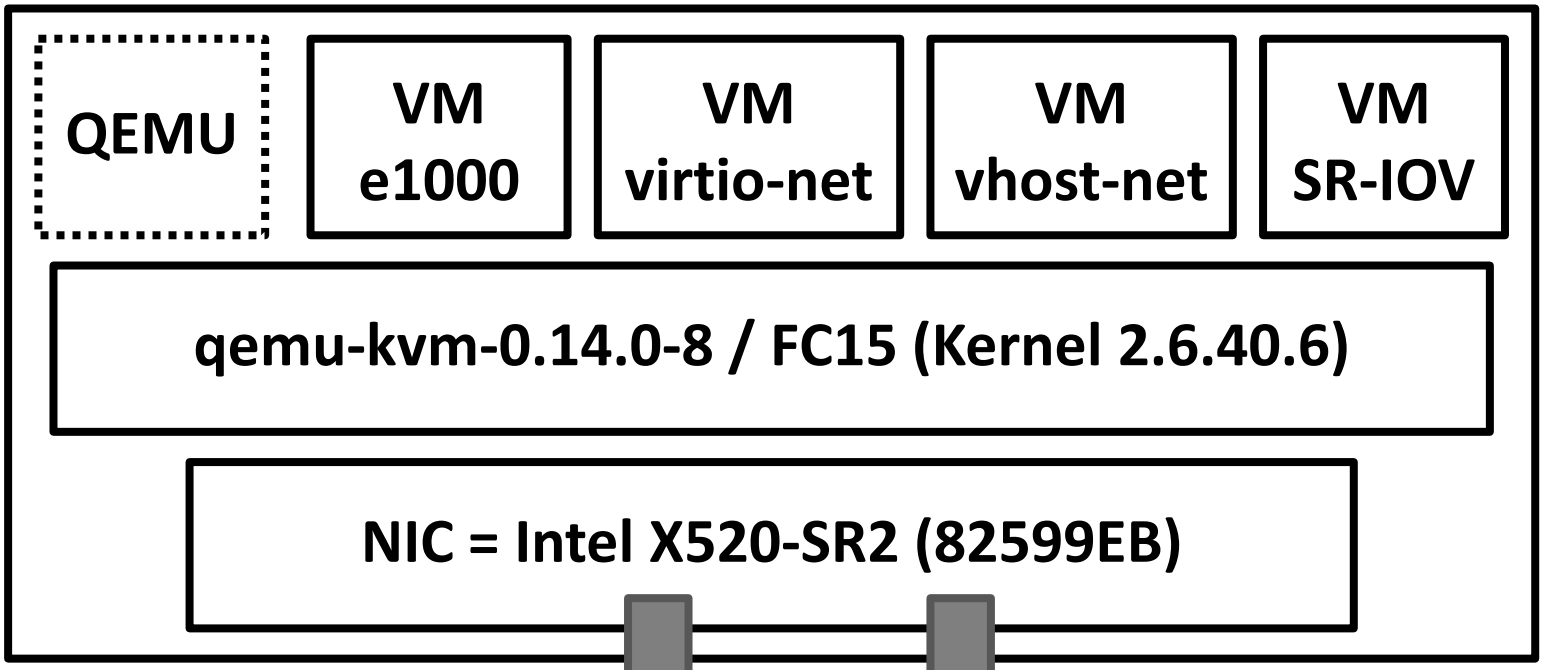
仮想化環境でのパケット転送
- 性能測定結果＆考察 -

海老澤 健太郎 @ パラレルス株式会社

2011/12/01

性能測定・構成

DUT (Device Under Test) – Dell PowerEdge R410
CPU x 2 : Xeon L5520 @ 2.27GHz



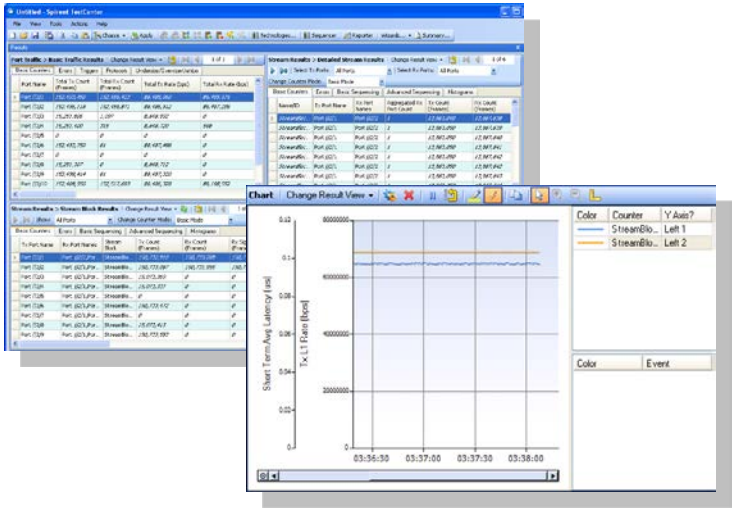
10GBase-SR x 2 port
uni- & bi-directional

機材紹介：Spirent TestCenter

次世代IP負荷測定擬似エミュレーション テスター

http://www.toyo.co.jp/spirenttestcenter/

＋



提供：東陽テクニカ

- SPT-2U：2U 2スロットシャーシ
- 10Gモジュール（10GBase-SR）

株式会社 東陽テクニカ

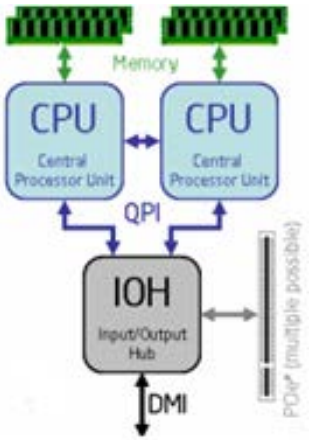
TESTED WITH SPIRENT

機材紹介：Server + NIC

- Dell PowerEdge R410（1U / 2 ソケット）
  CPU x 2 : Xeon L5520 @ 2.27GHz
  Chipset : Intel 5500

http://www.dell.com/jp/business/p/poweredge-r410/pd
http://ark.intel.com/products/36784/Intel-5500-IO-Hub

提供：東京大学



Dual Port
PCIe v2.0 (5.0GT/s)
SR-IOV

Networking Specifications	
# of Ports	Dual
Interface Type	PCIe v2.0 (5.0GT/s)
Temperature Range	0-70oC
Intel® Virtualization Technology for Connectivity (VT-c)	VMDq, SR-IOV
NC Sideband Interface	Yes
Fiber Channel over Ethernet	Yes
Jumbo Frames Supported	Yes
MACsec IEEE 802.1 AE	Yes
Supported Under vPro	No
Time Sync Protocol Indicator	Yes

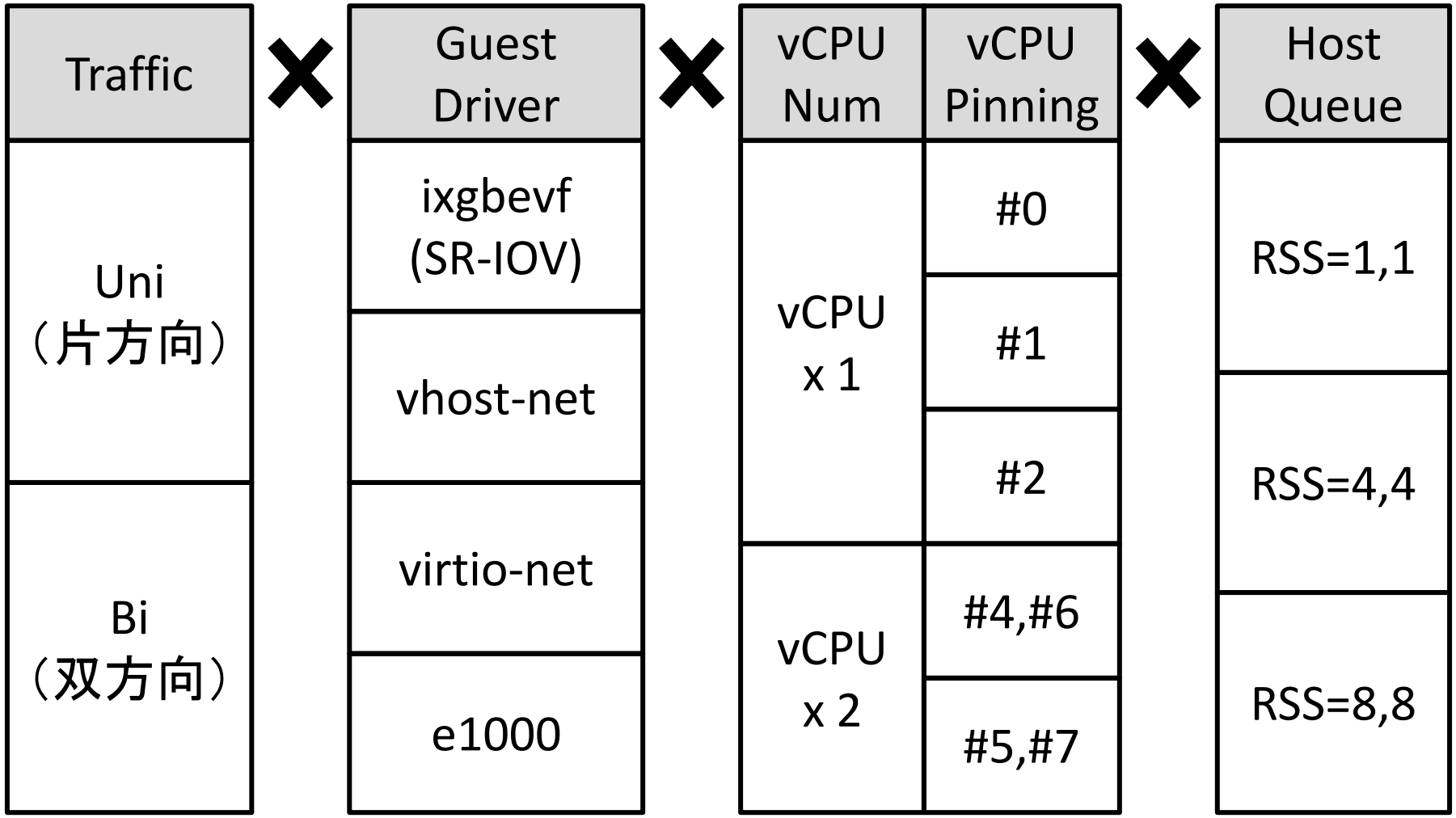
- INTEL® Ethernet Server Adapter : X520-SR2
  Ethernet controller: 82599EB

http://ark.intel.com/products/41282/Intel-82599ES-10-Gigabit-Ethernet-Controller
http://ark.intel.com/products/39774/Intel-Ethernet-Server-Adapter-X520-SR2

提供：さくらインターネット

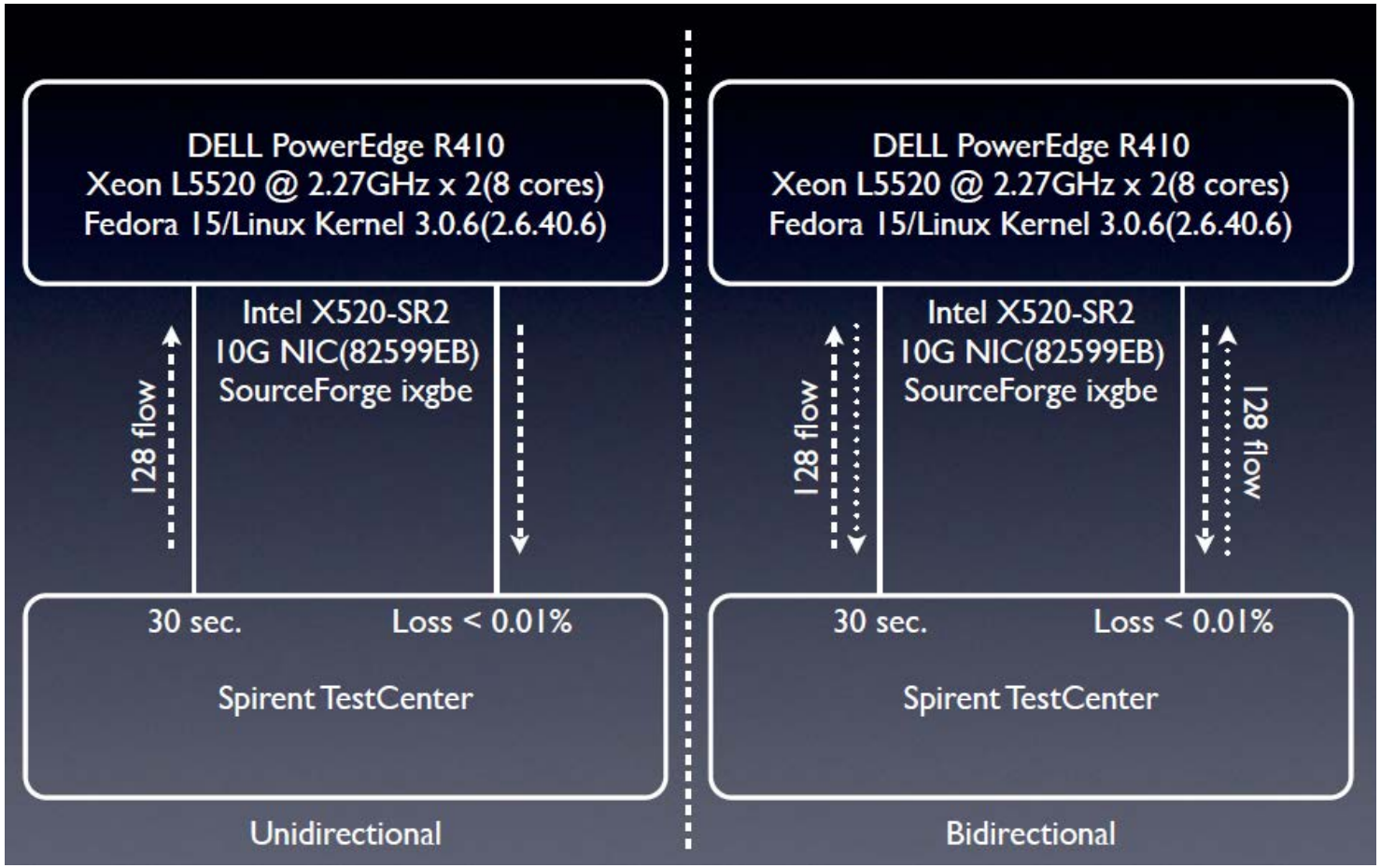
SAKURA Internet

計測パターン

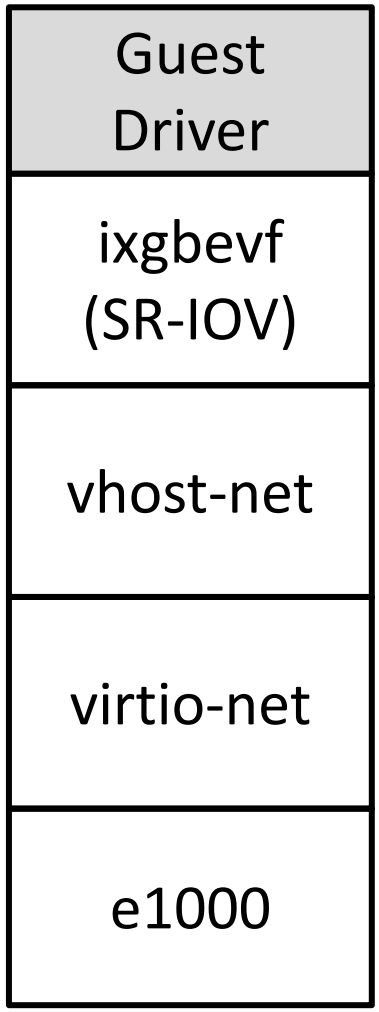

Traffic
Guest Driver
vCPU Num
vCPU Pinning
Host Queue
Uni（片方向）
Bi（双方向）
ixgbevf (SR-IOV)
vhost-net
virtio-net
e1000
vCPU x 1
vCPU x 2
#0
#1
#2
#4,#6
#5,#7
RSS=1,1
RSS=4,4
RSS=8,8

測定パターン：Traffic

Traffic
Uni （片方向）
Bi （双方向）



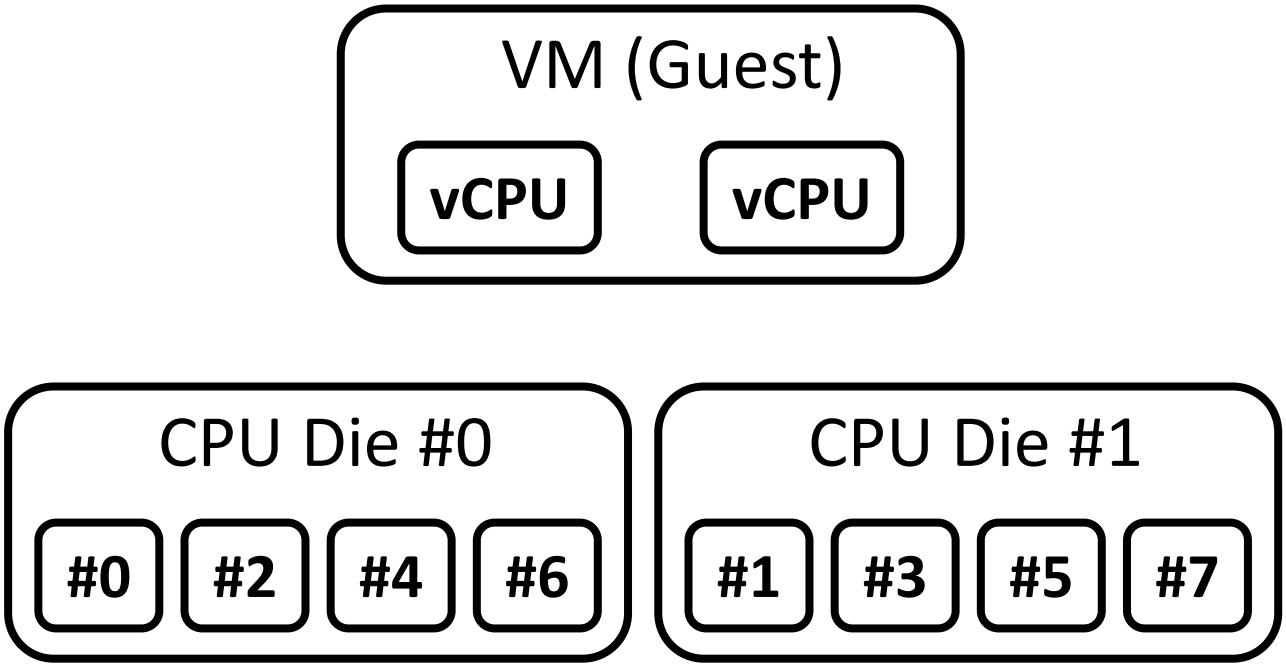
計測パターン：Guest Driver



使用したGuest Driver
※ 浅田さん資料参照

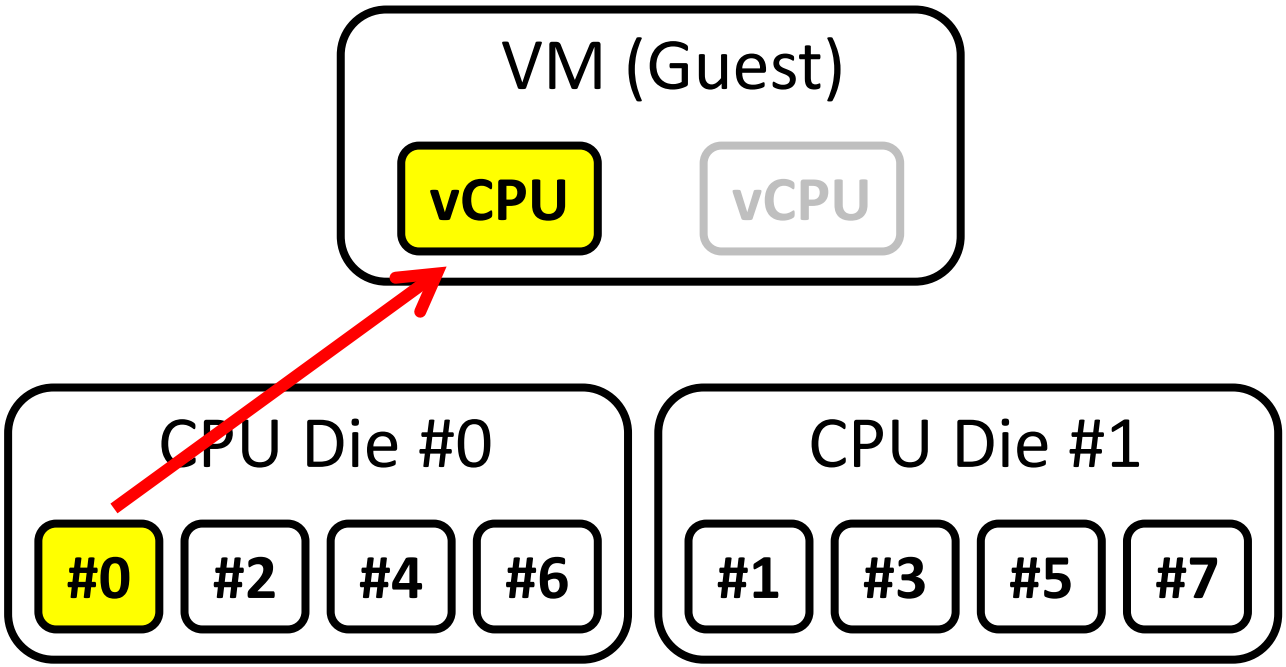
計測パターン：vCPU Num, Pinning

vCPU Num	vCPU Pinning
vCPU x 1	#0
	#1
	#2
vCPU x 2	#4,#6
	#5,#7



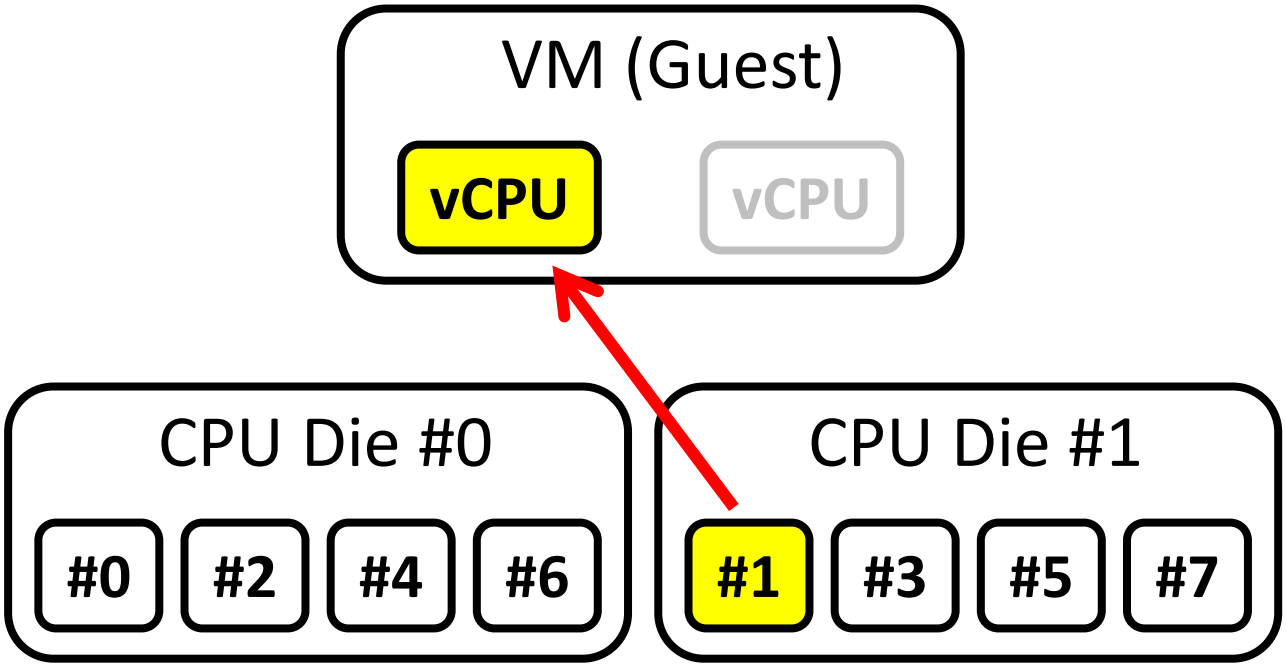
計測パターン：vCPU Num, Pinning

vCPU Num	vCPU Pinning
vCPU x 1	#0
	#1
	#2
vCPU x 2	#4,#6
	#5,#7



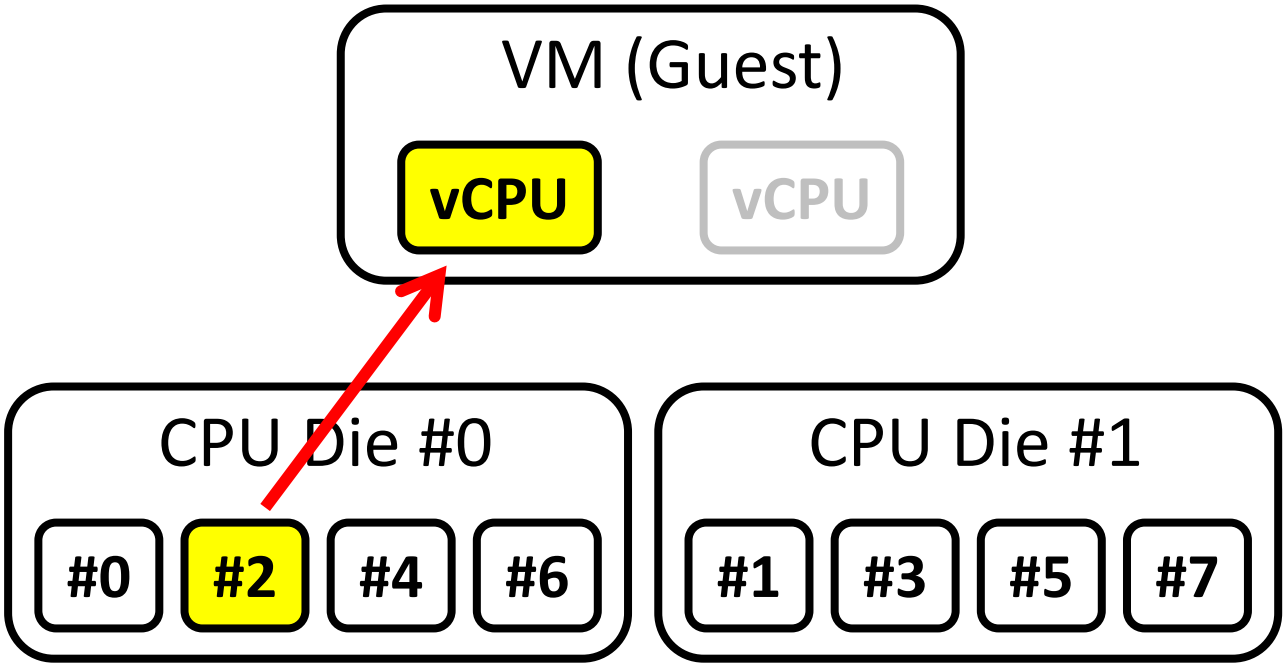
計測パターン：vCPU Num, Pinning

vCPU Num	vCPU Pinning
vCPU x 1	#0
	#1
	#2
vCPU x 2	#4,#6
	#5,#7



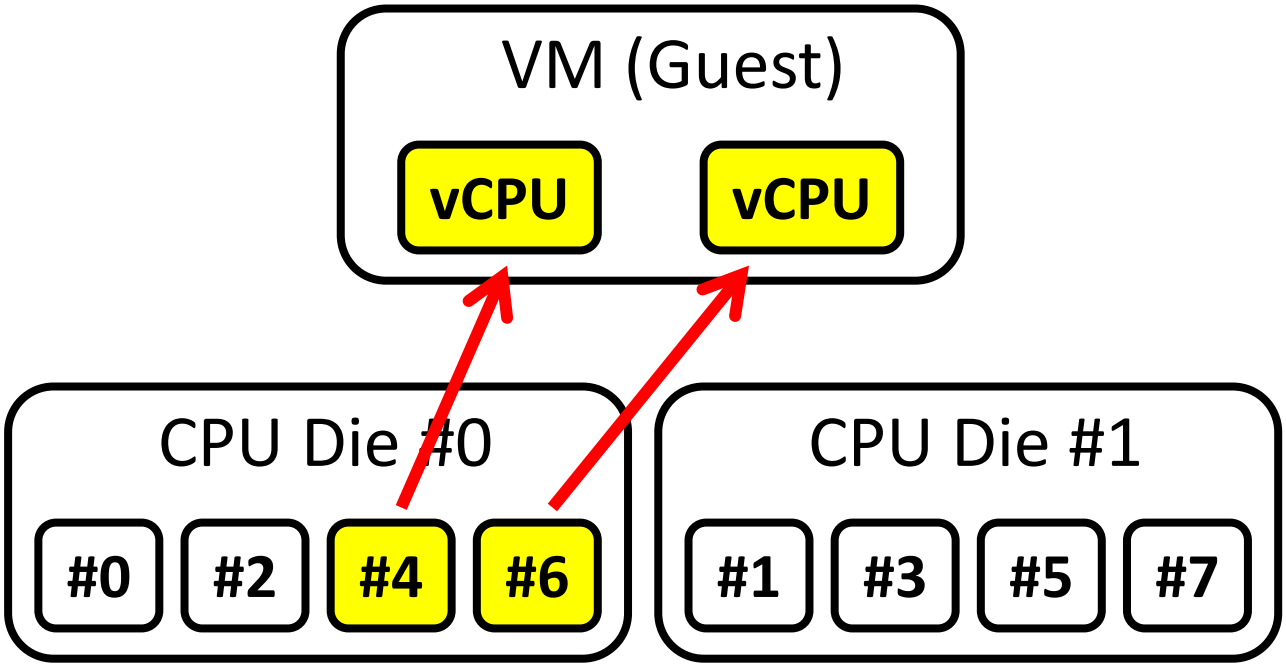
計測パターン：vCPU Num, Pinning

vCPU Num	vCPU Pinning
vCPU x 1	#0
	#1
	#2
vCPU x 2	#4,#6
	#5,#7



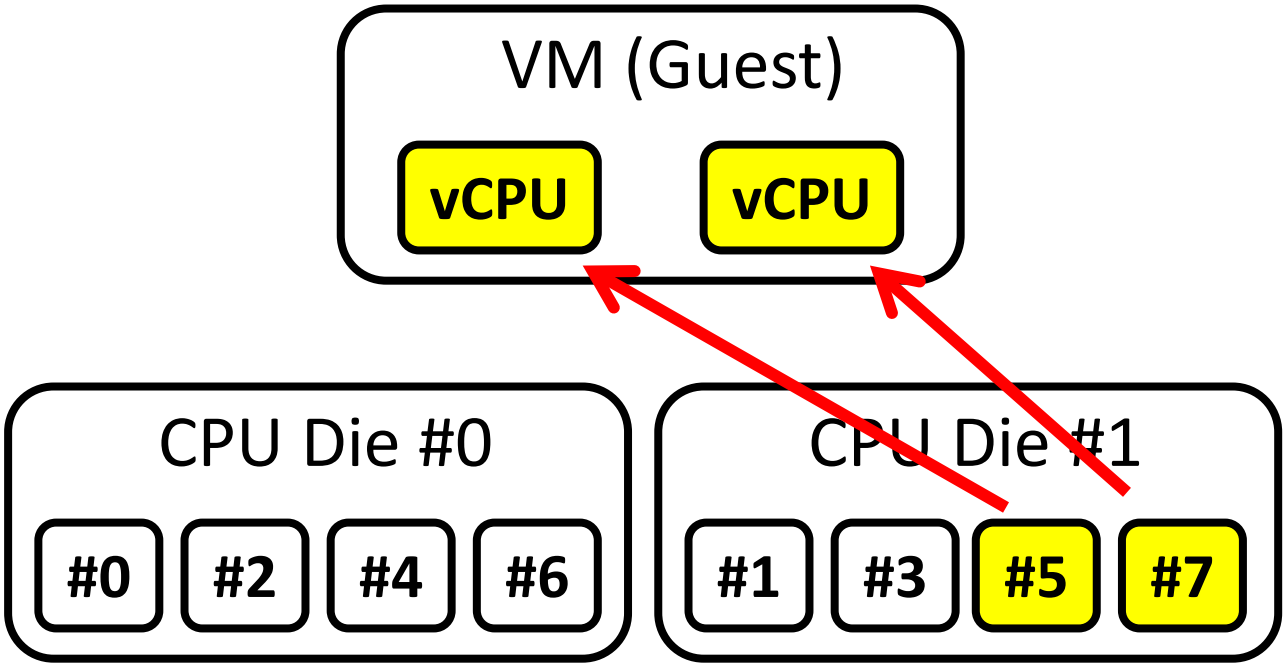
計測パターン：vCPU Num, Pinning

vCPU Num	vCPU Pinning
vCPU x 1	#0
	#1
	#2
vCPU x 2	#4,#6
	#5,#7

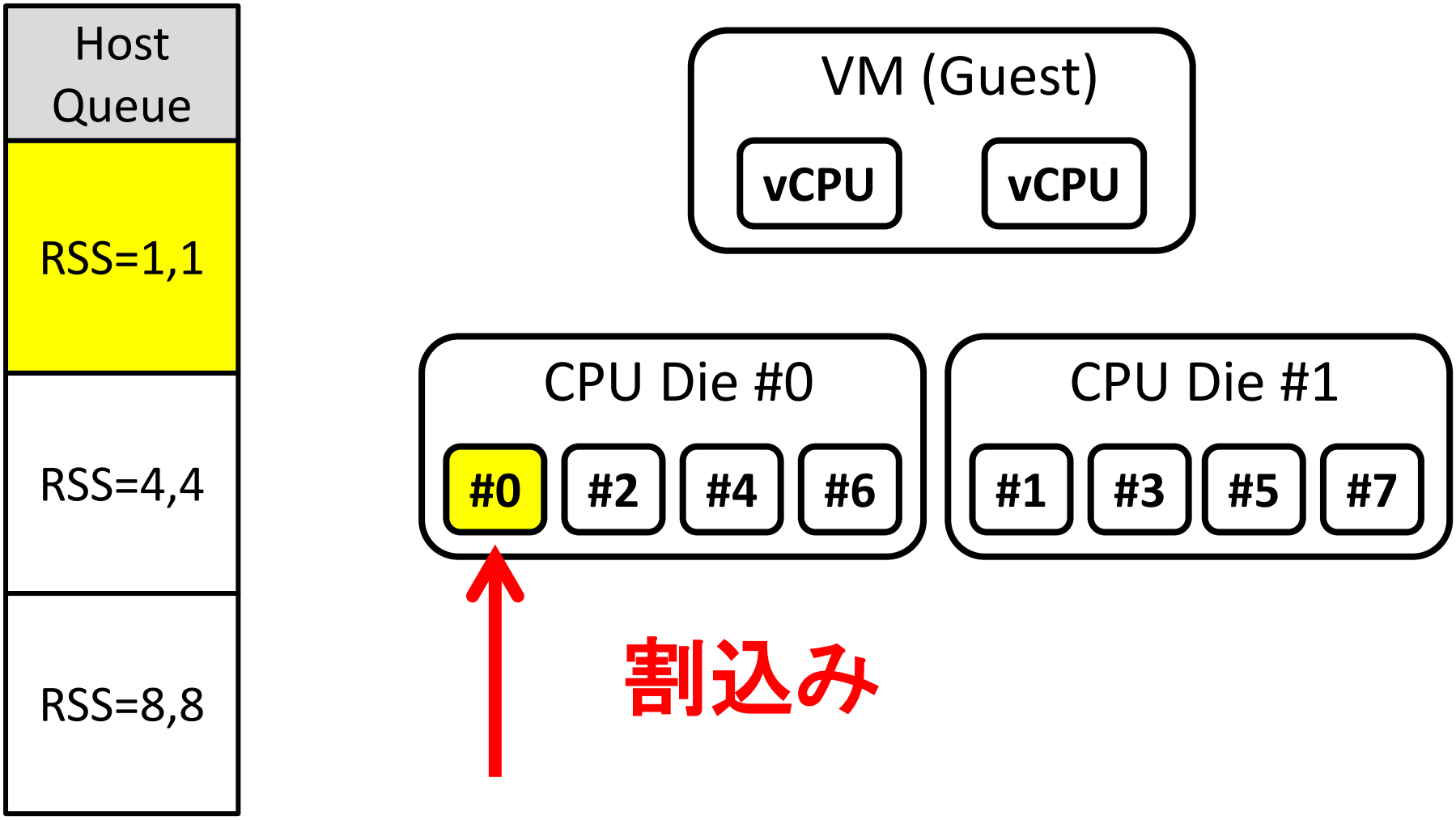


計測パターン：vCPU Num, Pinning

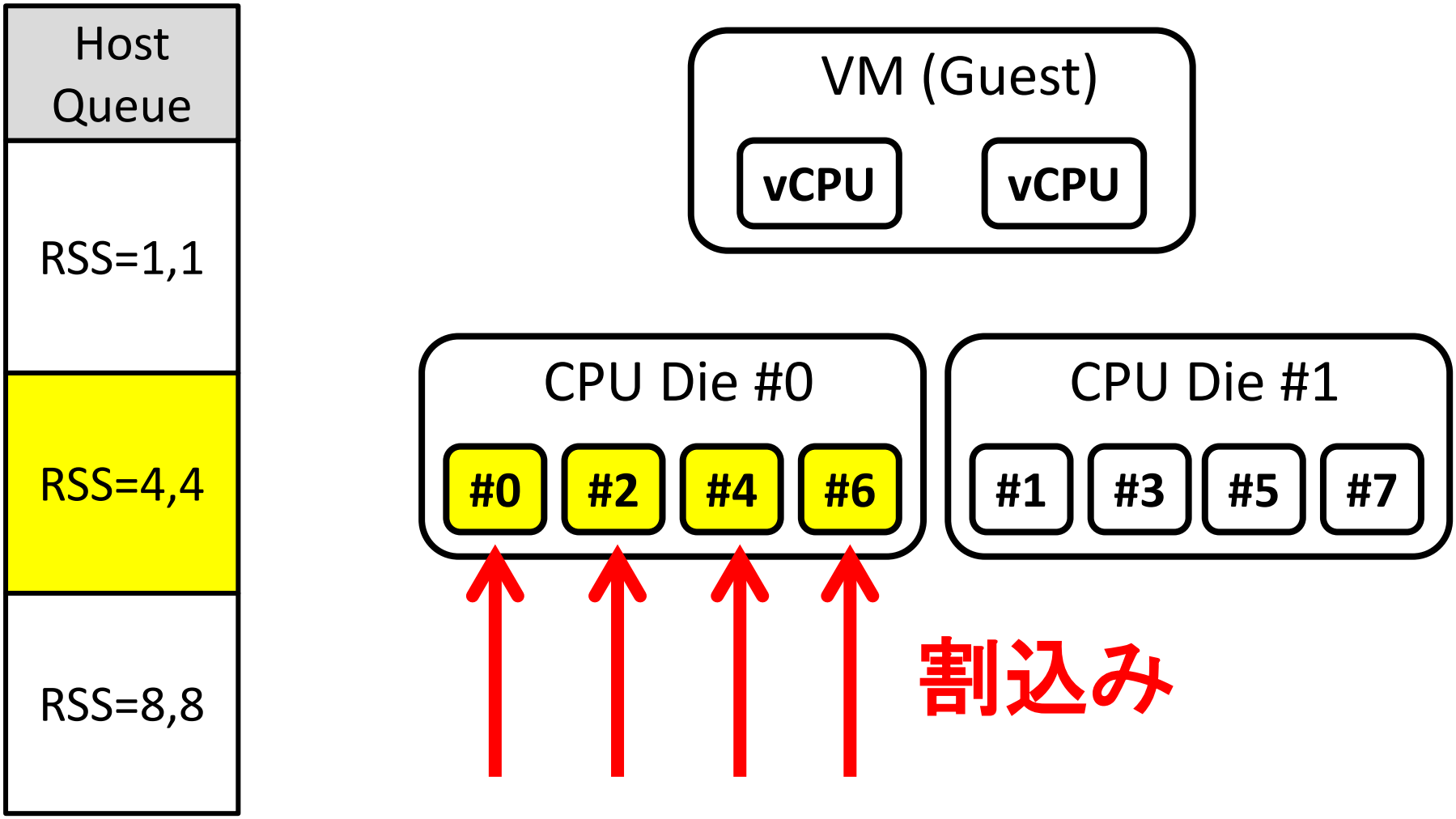
vCPU Num	vCPU Pinning
vCPU x 1	#0
	#1
	#2
vCPU x 2	#4,#6
	#5,#7



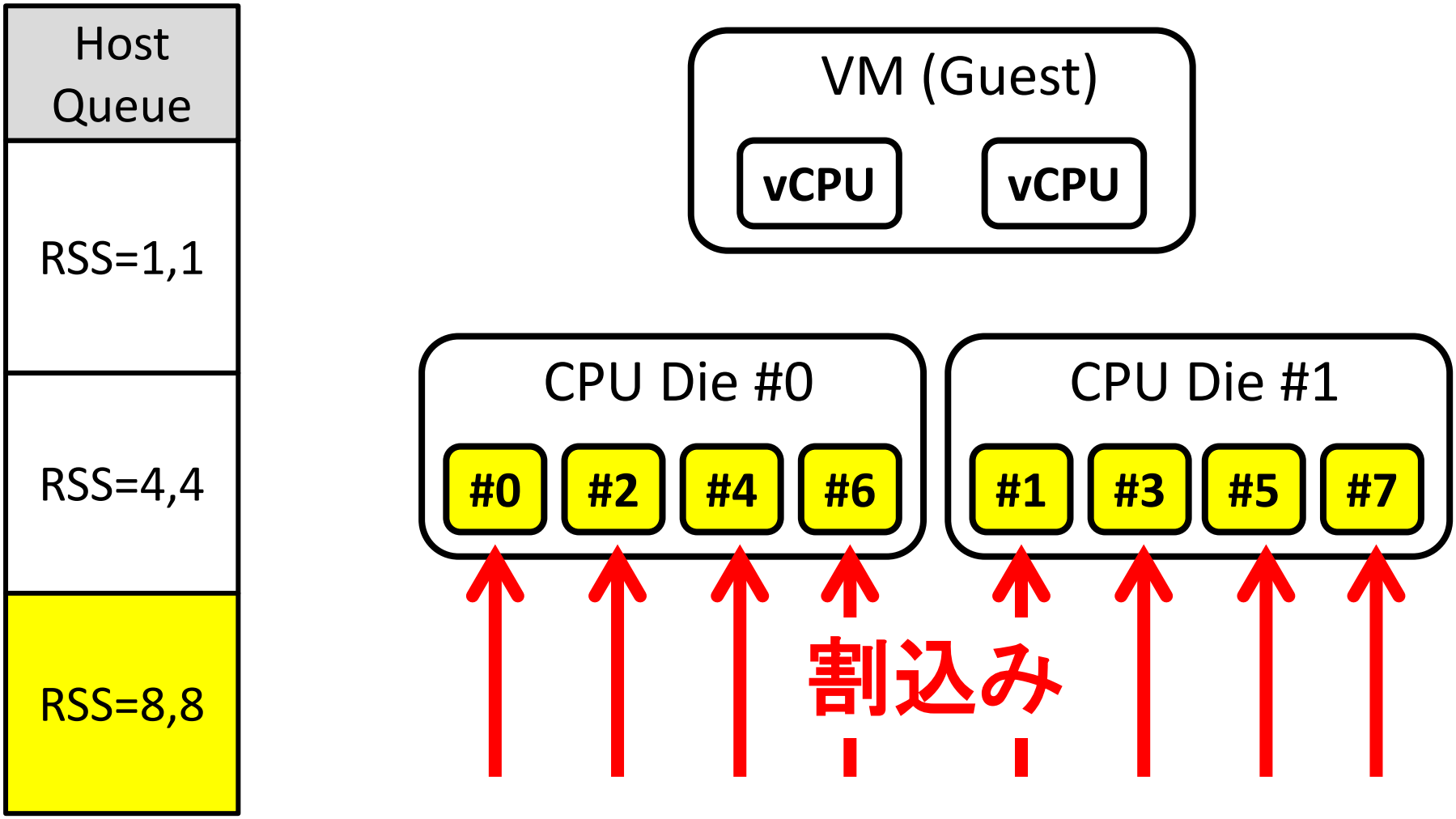
計測パターン：Host Queue



計測パターン：Host Queue

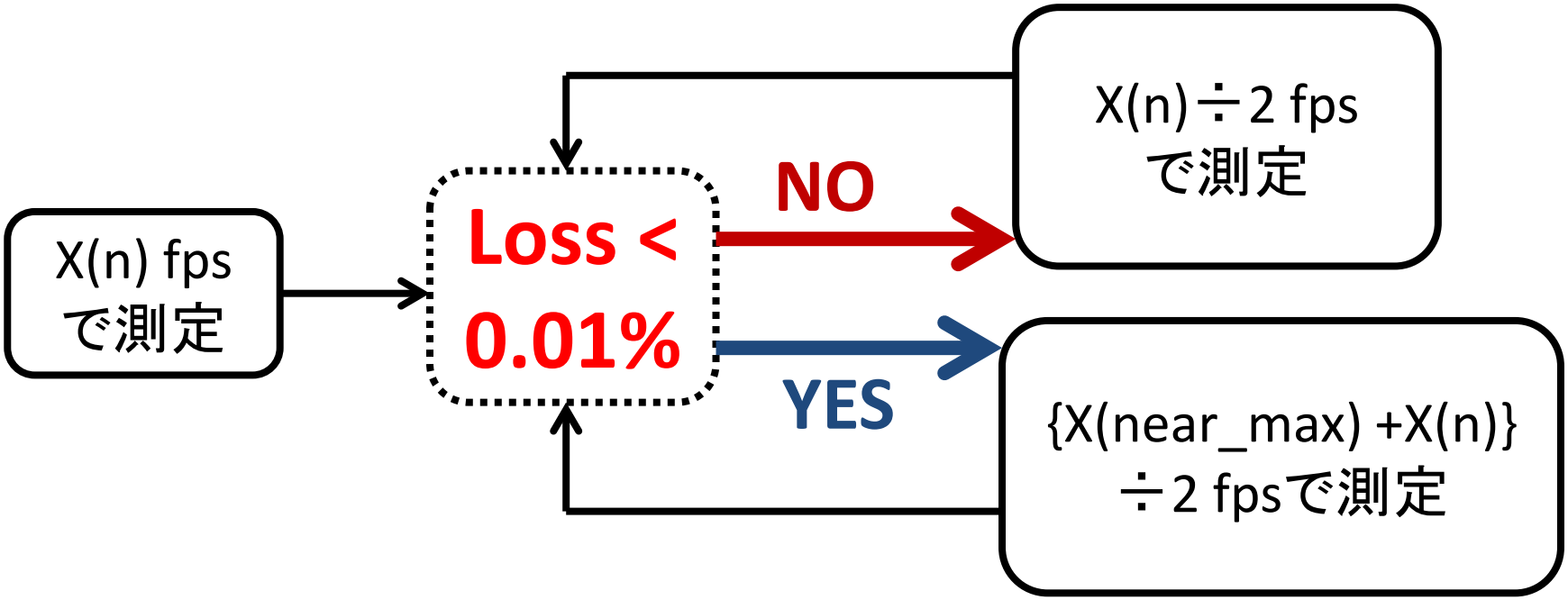


計測パターン：Host Queue



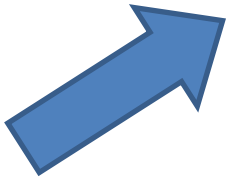
測定方法及び誤差に関して

- Spirent Test Center - Command Sequencer
  - Load Type = Binary
  - Acceptable Frame Loss < 0.01%



Binary + Acceptable Frame Loss 測定の特徴

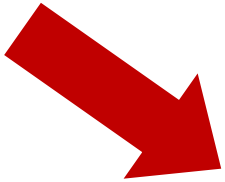
パケットロスのほぼ無い状況での最大性能を測定可能



Good

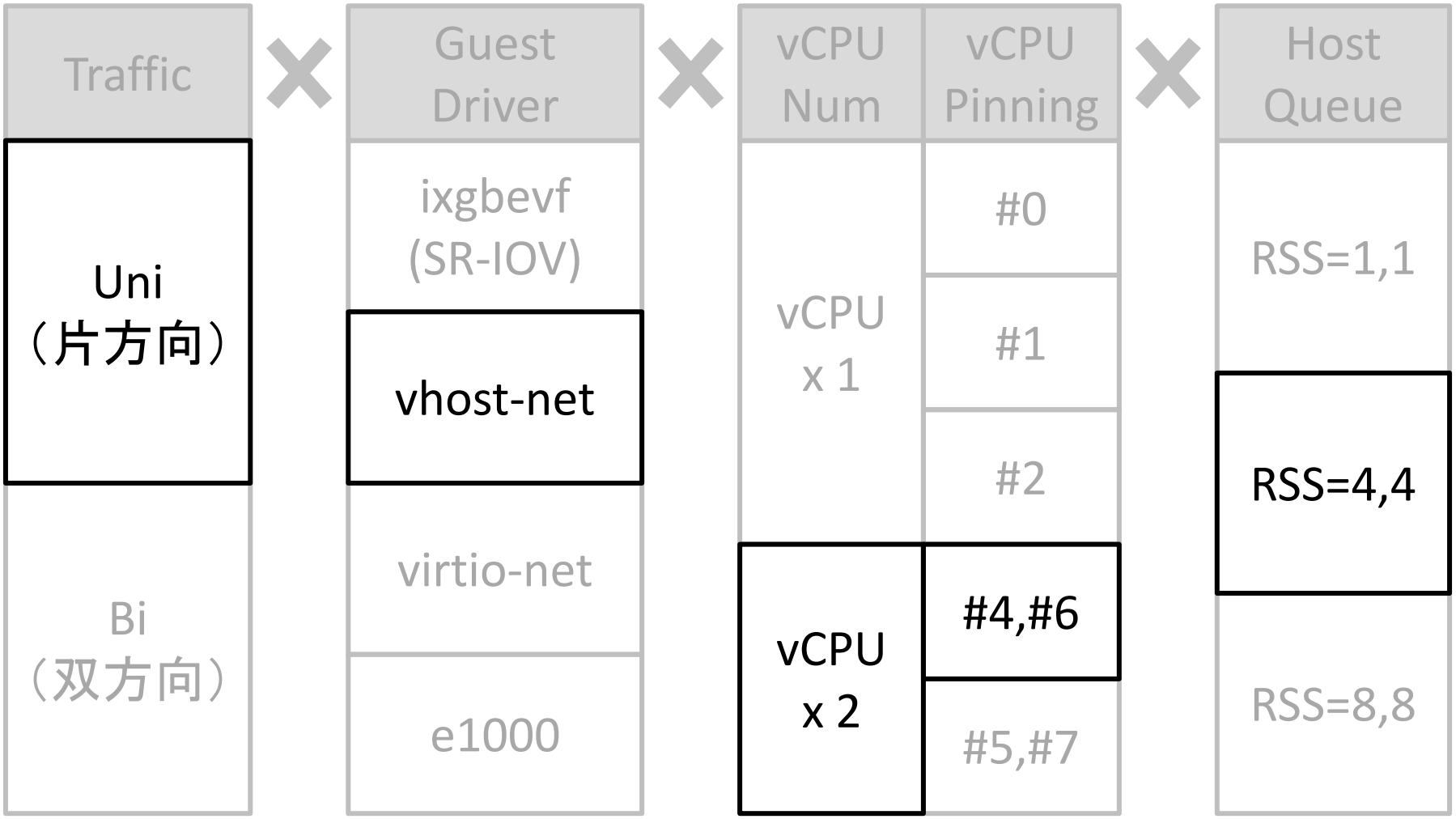
バースト＆受信 fps 測定で発生しがちな高負荷時の性能劣化により、最大性能が低く測定される事がない

Bad

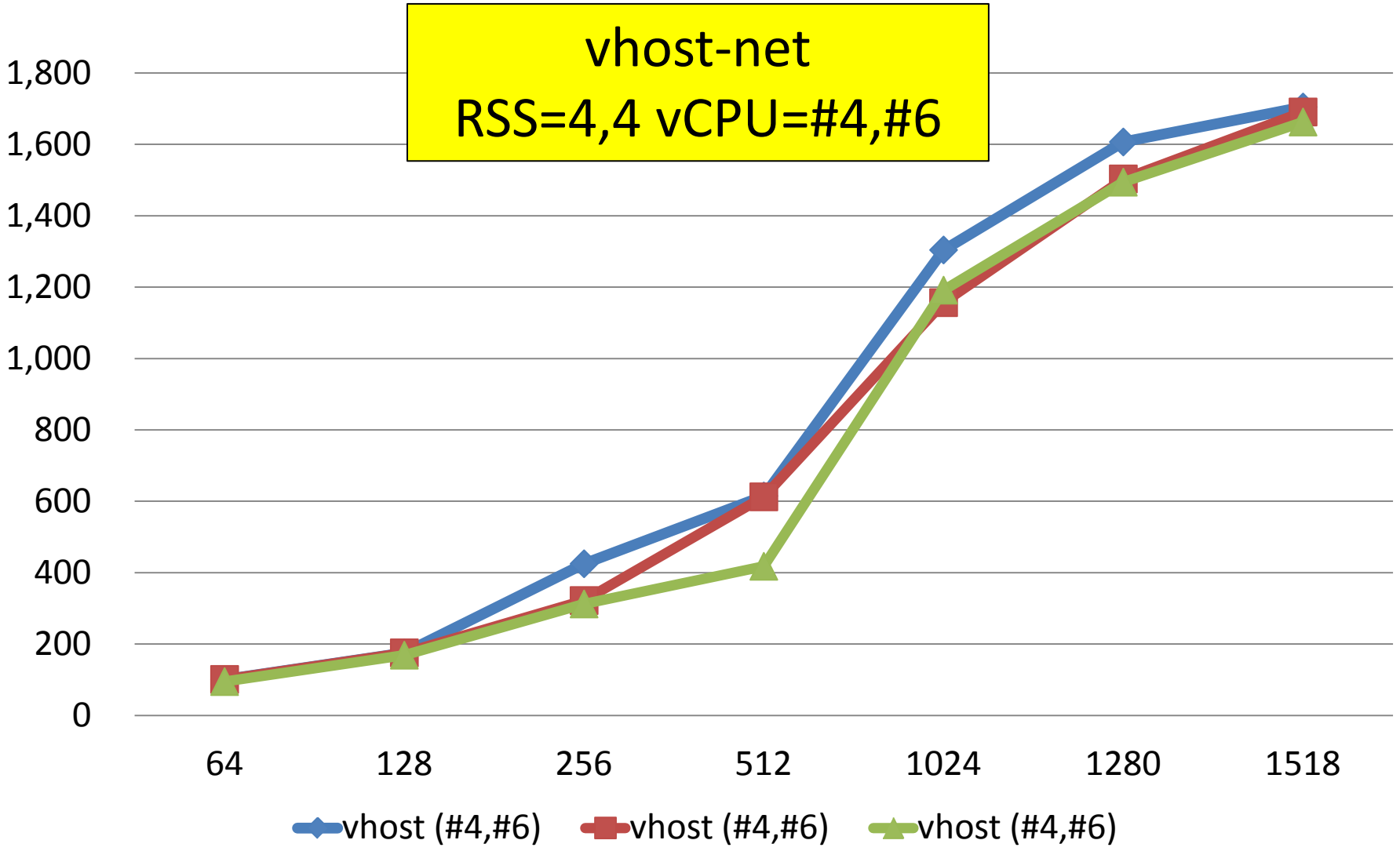


トラフィック開始直後にロスが発生する等、実環境で無視可能な事象により実際の性能より低く測定される場合がある

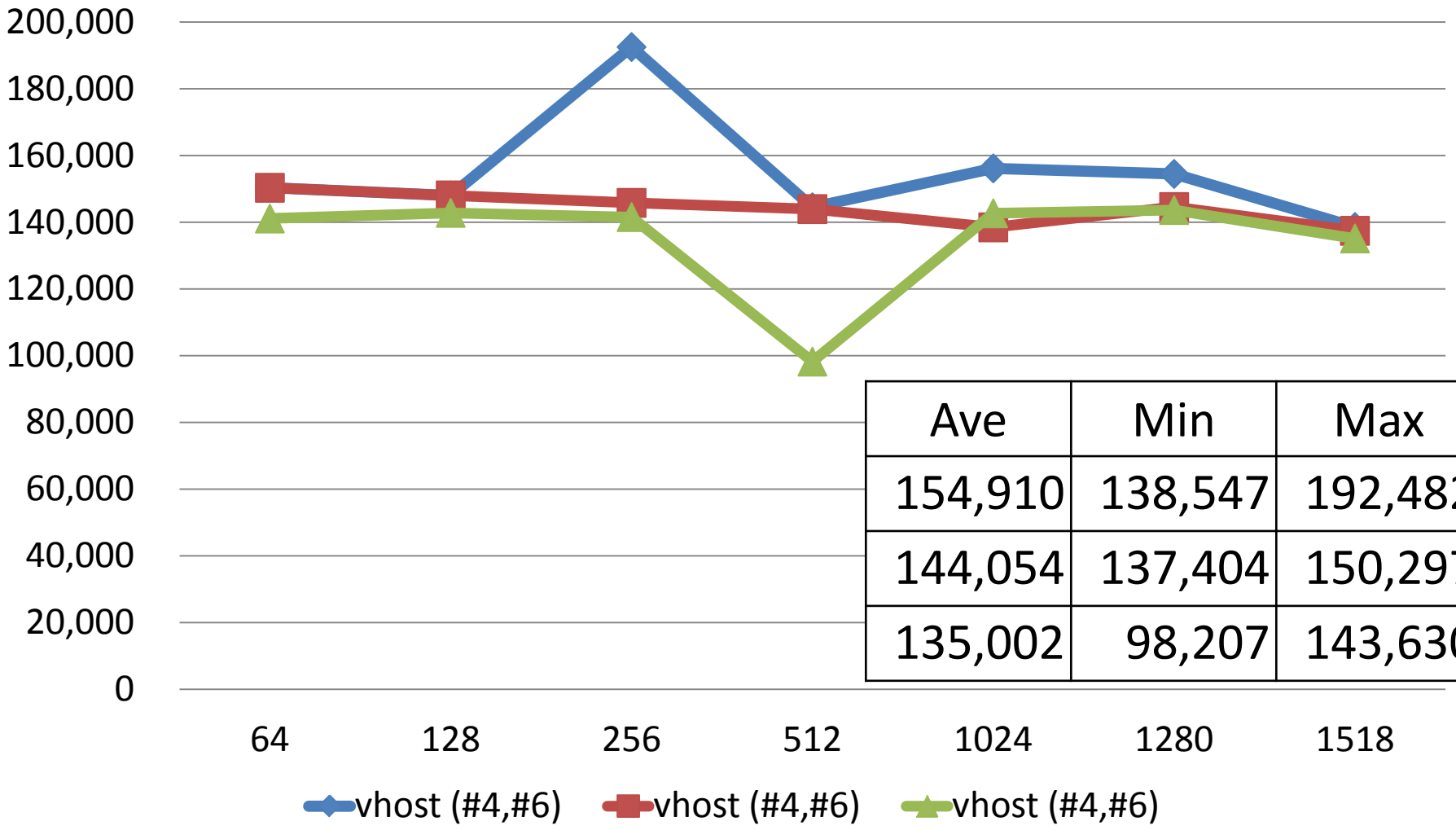
測定方法及び誤差に関して



測定方法及び誤差に関して:同一条件x3試行



測定方法及び誤差に関して:同一条件x3試行



Ave	Min	Max
154,910	138,547	192,482
144,054	137,404	150,297
135,002	98,207	143,630

CPU負荷率とパケットロス

```
top - 07:32:22 up 10 min,  1 user,  load average: 0.00, 0.00, 0.00
Tasks:  54 total,   1 running,  53 sleeping,   0 stopped,   0 zombie
Cpu0  :  0.0%us,  0.0%sy,  0.0%ni, 38.2%id,  0.0%wa, 47.0%hi, 14.8%si,  0.0%st
Cpu1  :  0.0%us,  0.3%sy,  0.0%ni, 37.3%id,  0.0%wa, 46.9%hi, 15.5%si,  0.0%st
Mem:   1023332k total,    63300k used,   960032k free,     5944k buffers
Swap:  2031612k total,        0k used,  2031612k free,    18088k cached
```

30%~40% idle でもパケットロス発生

vhost-net : vCPUx2 @ 120fps (loss 2%)

```
top - 07:41:37 up  7:36,  3 users,  load average: 0.07, 0.12, 0.16
Tasks: 123 total,   3 running, 118 sleeping,   0 stopped,   2 zombie
Cpu0  :  0.0%us, 17.3%sy,  0.0%ni, 18.6%id,  0.0%wa,  4.0%hi, 60.1%si,  0.0%st
Cpu1  :  1.0%us,  6.3%sy,  0.0%ni, 86.5%id,  0.0%wa,  0.7%hi,  5.6%si,  0.0%st
Cpu2  :  0.0%us, 27.2%sy,  0.0%ni, 43.7%id,  0.0%wa,  0.7%hi, 28.5%si,  0.0%st
Cpu3  :  1.3%us,  1.7%sy,  0.0%ni, 96.6%id,  0.0%wa,  0.0%hi,  0.4%si,  0.0%st
Cpu4  : 42.5%us, 24.9%sy,  0.0%ni, 30.9%id,  0.0%wa,  1.7%hi,  0.0%si,  0.0%st
Cpu5  :  0.7%us,  0.7%sy,  0.0%ni, 98.7%id,  0.0%wa,  0.0%hi,  0.0%si,  0.0%st
Cpu6  : 44.3%us, 24.7%sy,  0.0%ni, 29.3%id,  0.0%wa,  1.7%hi,  0.0%si,  0.0%st
Cpu7  :  0.0%us,  0.7%sy,  0.0%ni, 99.3%id,  0.0%wa,  0.0%hi,  0.0%si,  0.0%st
Mem:  33005580k total,   984656k used, 32020924k free,    32280k buffers
Swap: 35160060k total,        0k used, 35160060k free,   271228k cached
```

パケット
ロス発生

≠

CPU（Core）
使用率100%

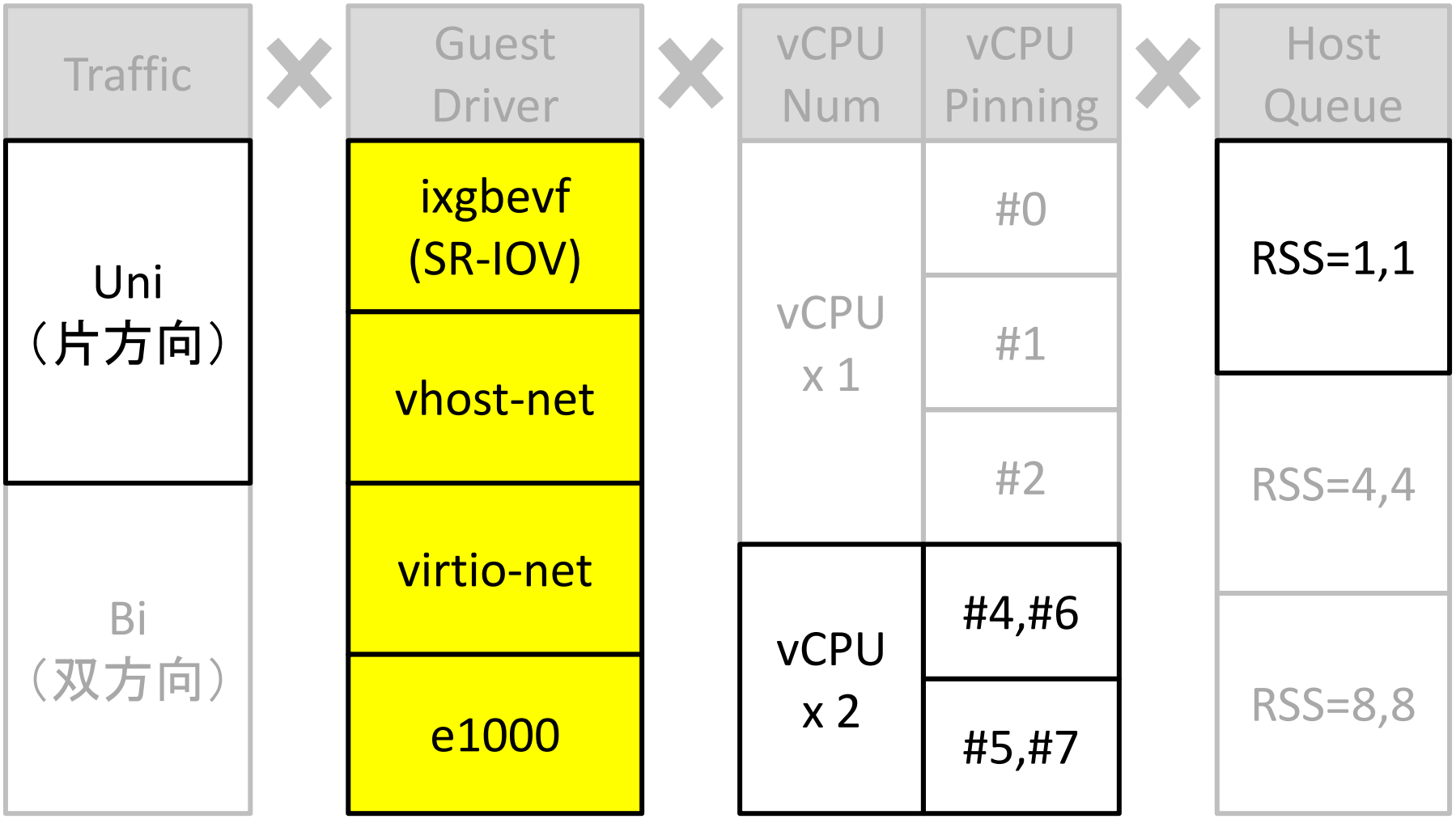
測定の目的は？？

？安定的な転送性能（HW Switch）？

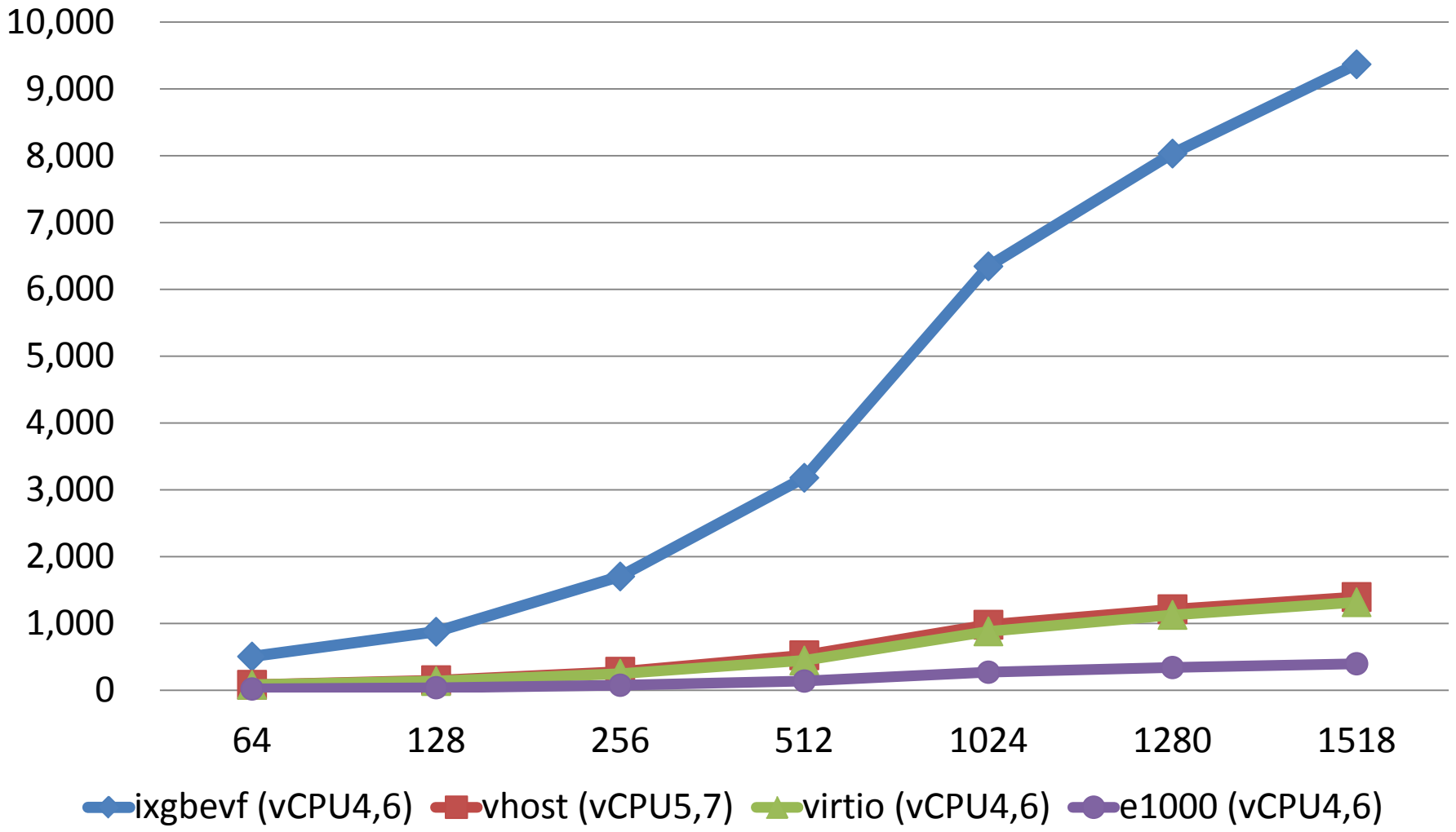
？バースト時のピーク？

？機器・技術の性能特性？

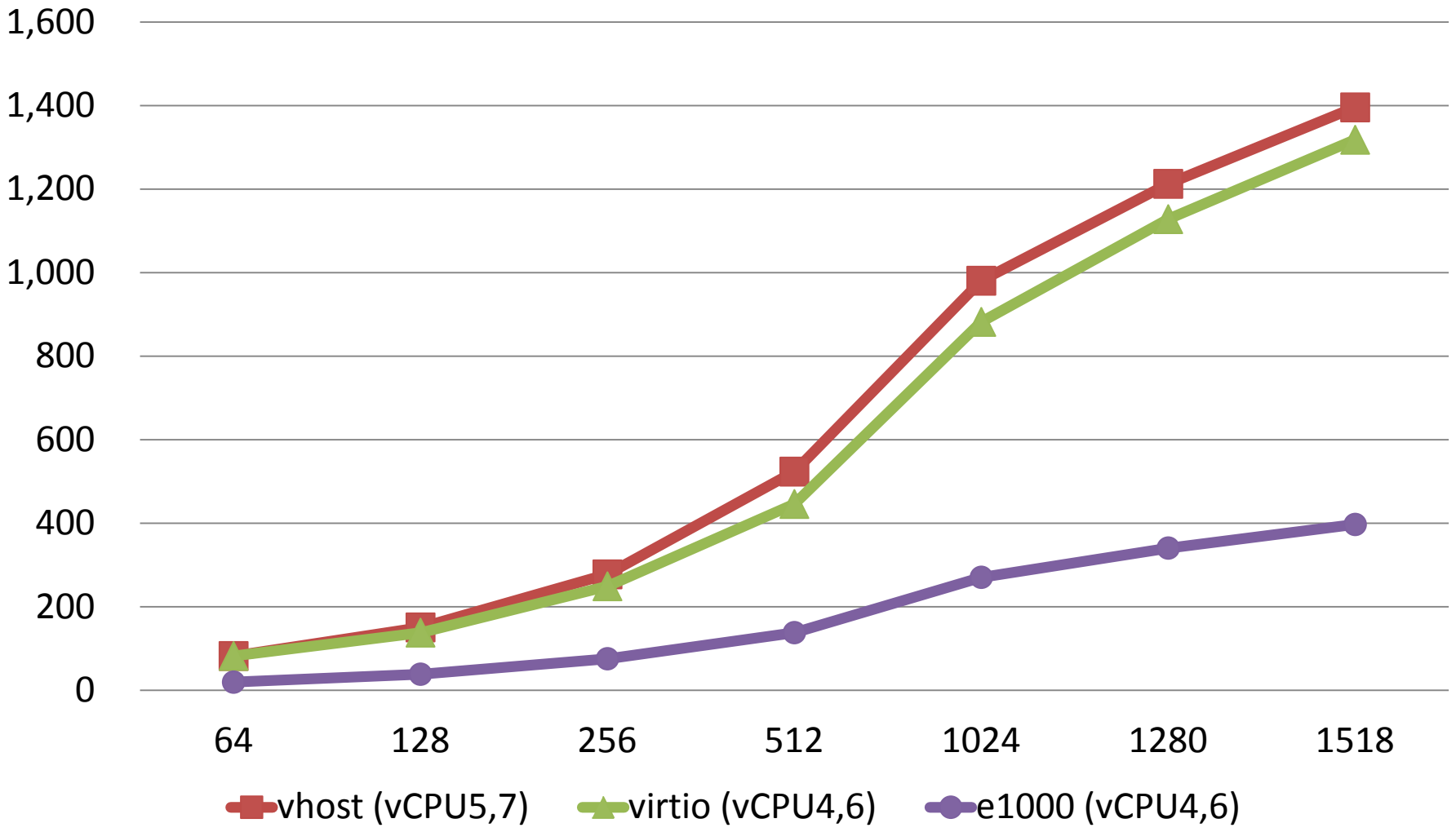
Host Driver による性能差



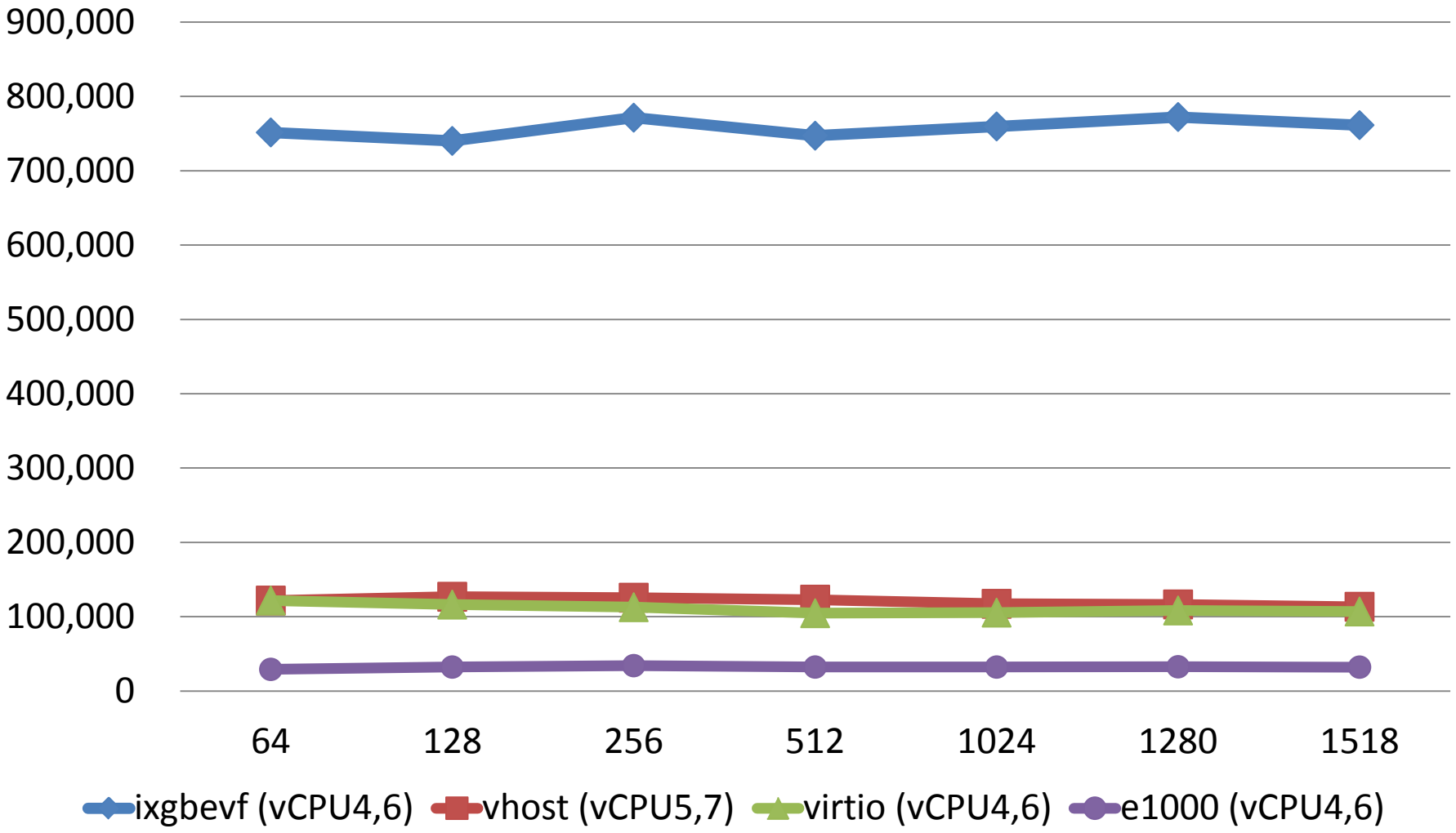
Host Driver による性能差 (Mbps)



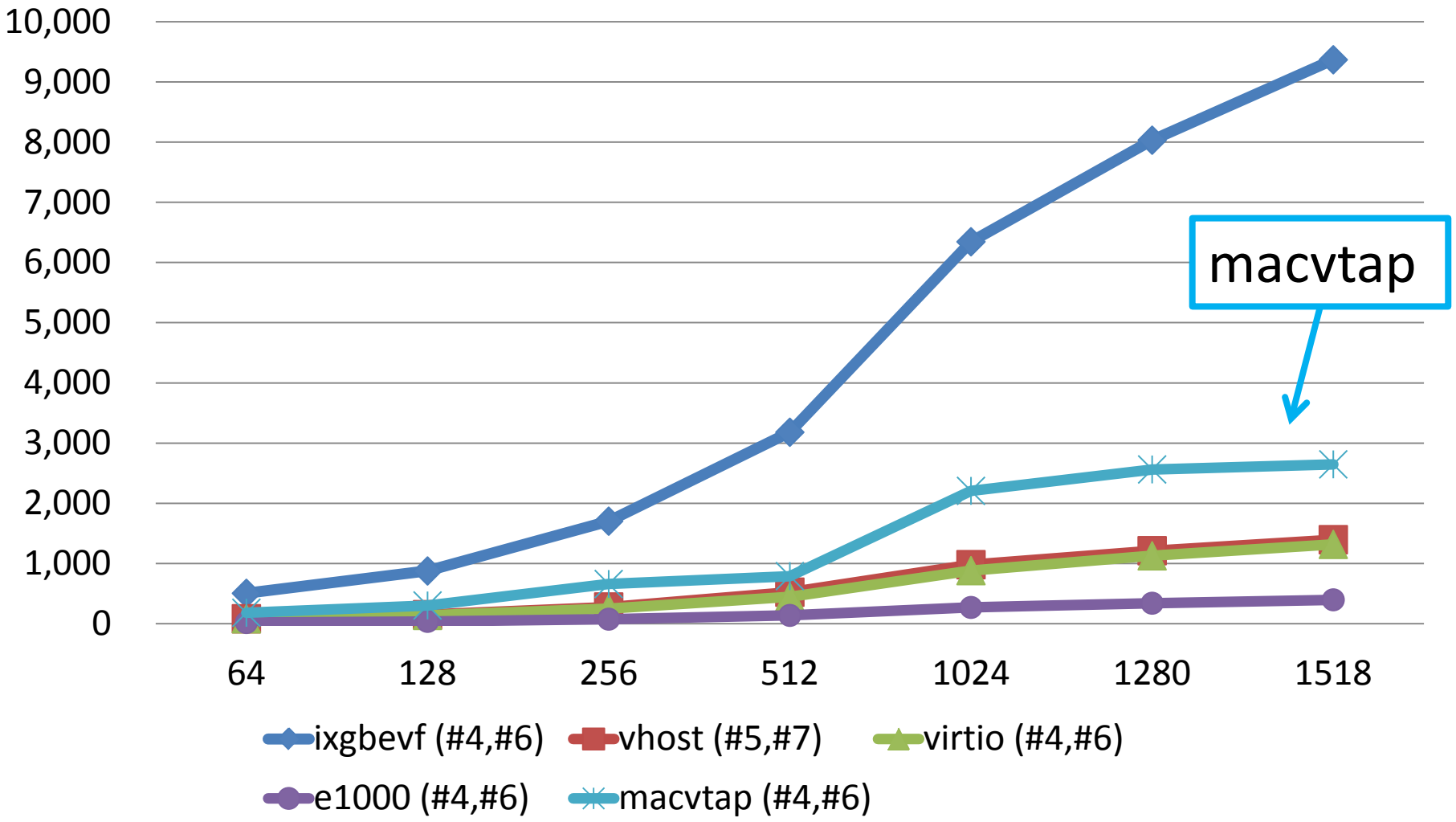
Host Driver による性能差 (Mbps)



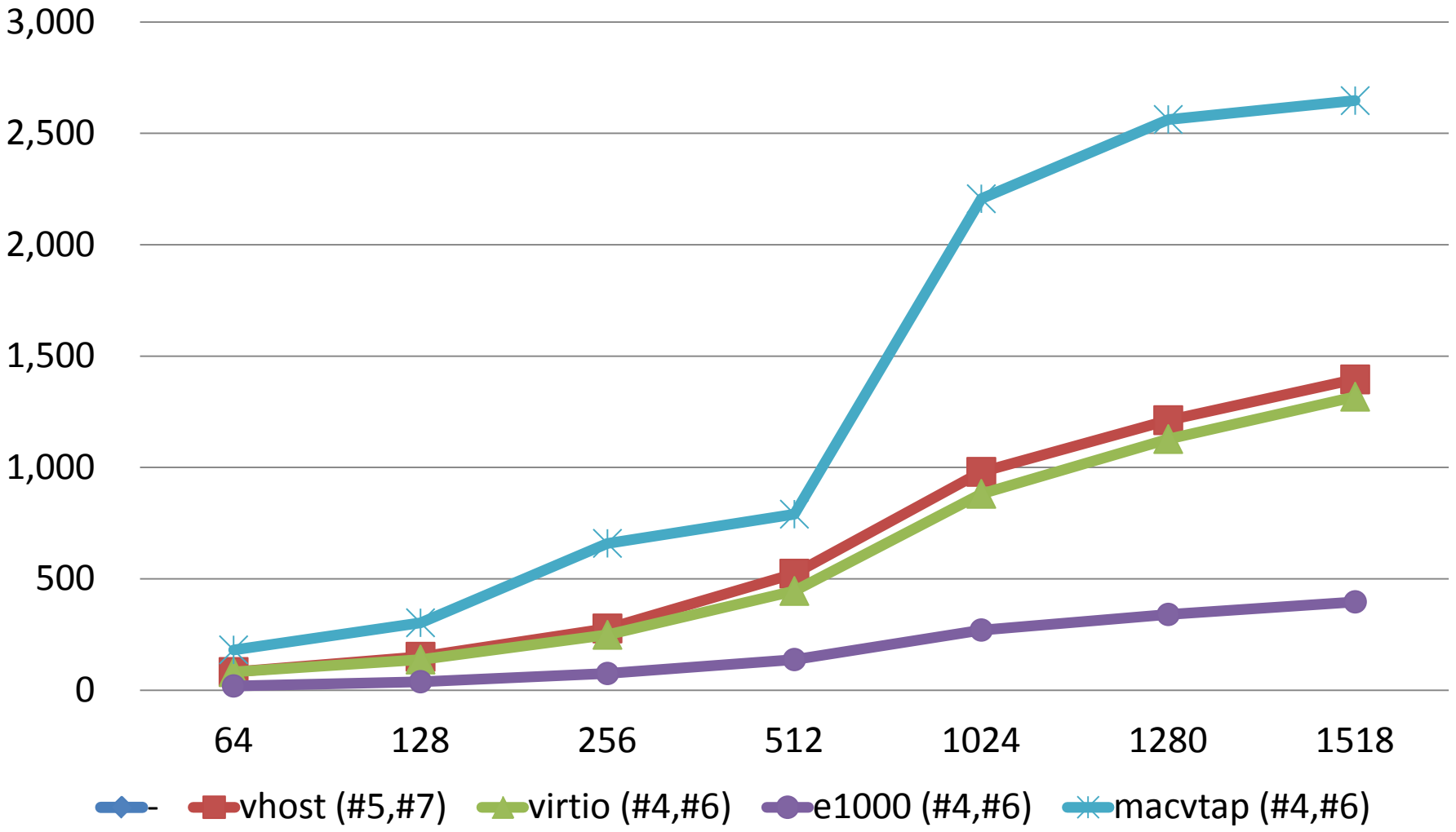
Host Driver による性能差 (pps)



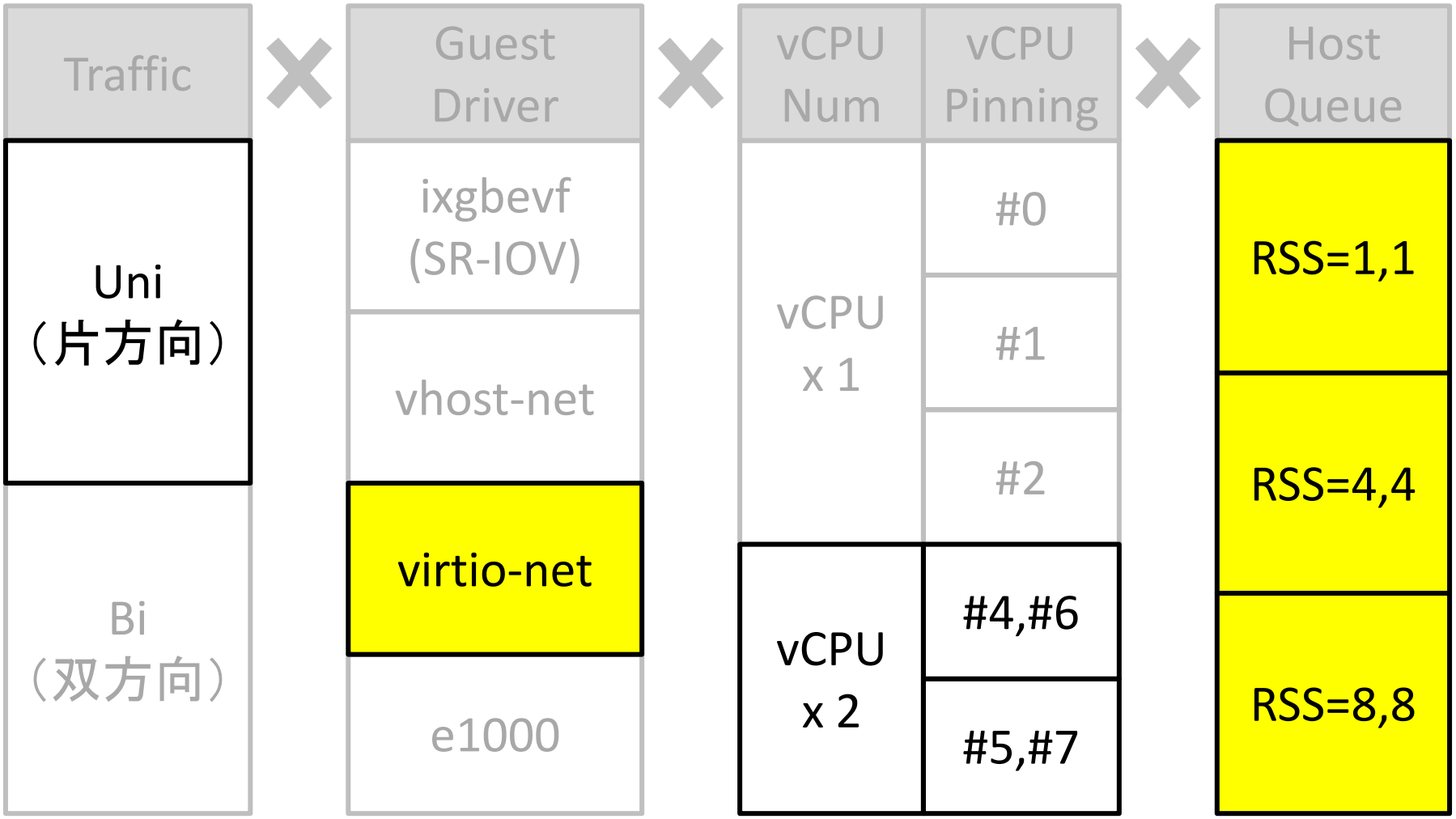
Host Driver による性能差 (Mbps)



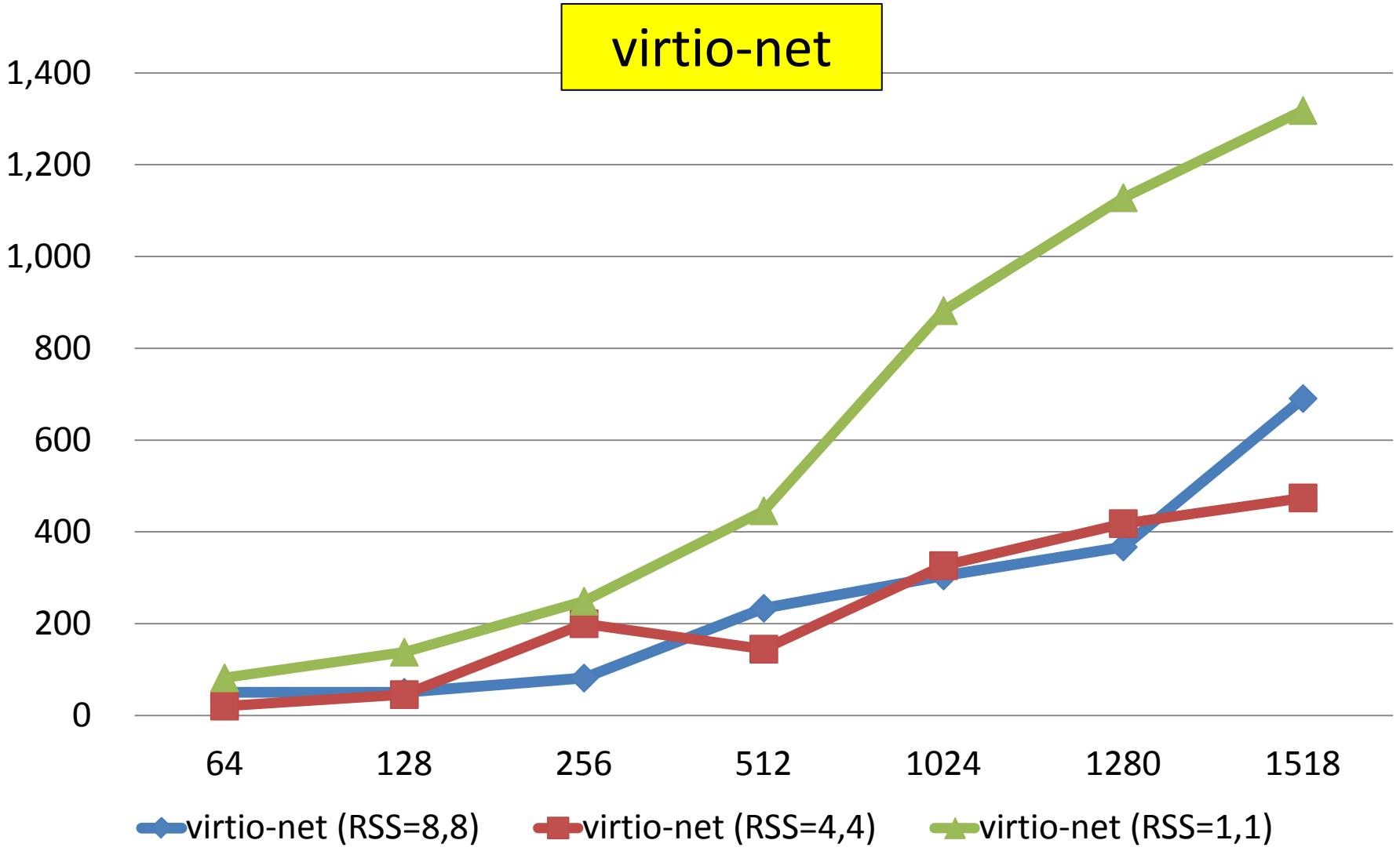
Host Driver による性能差 (Mbps)



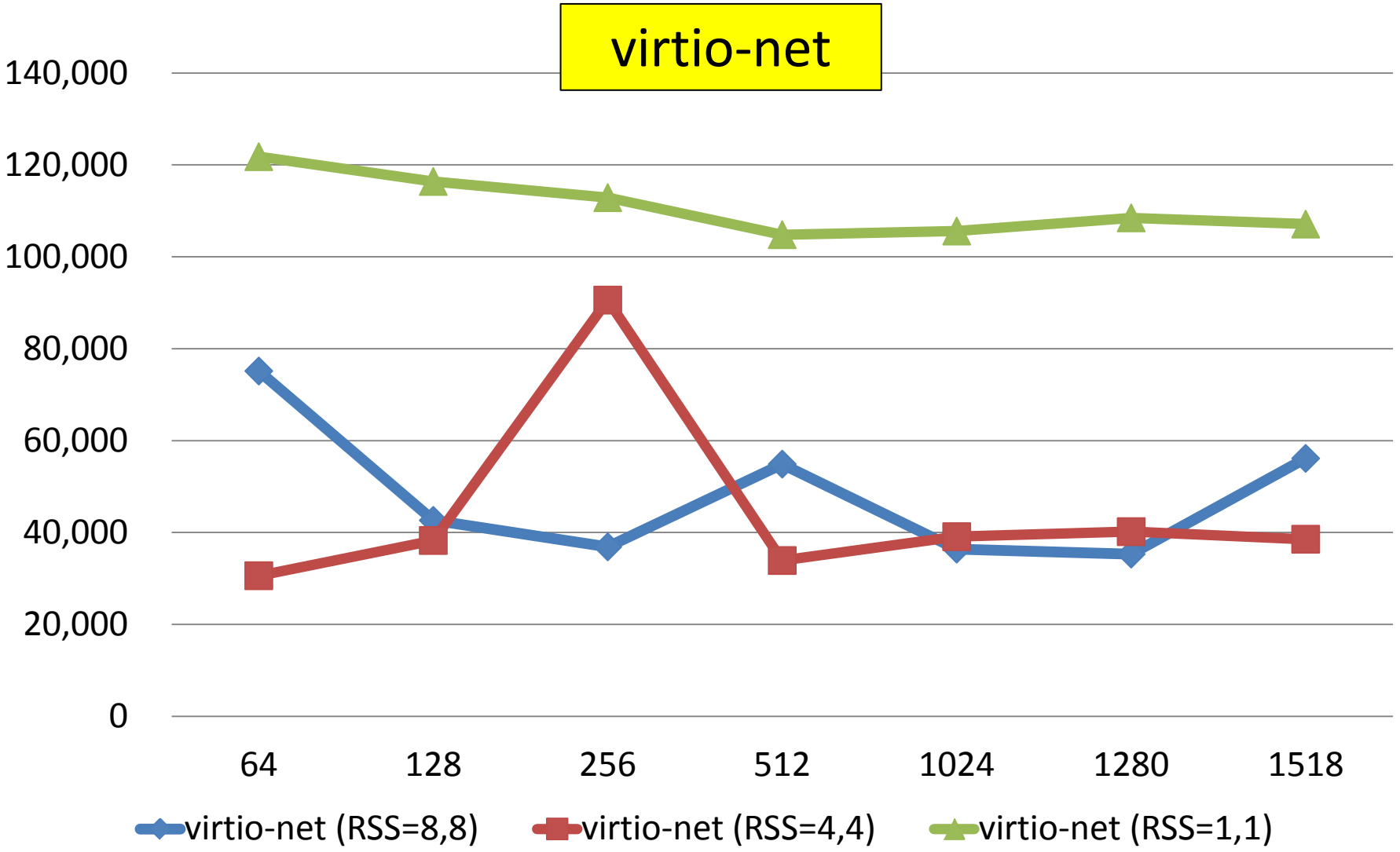
Host Queue（RSS）設定による性能差



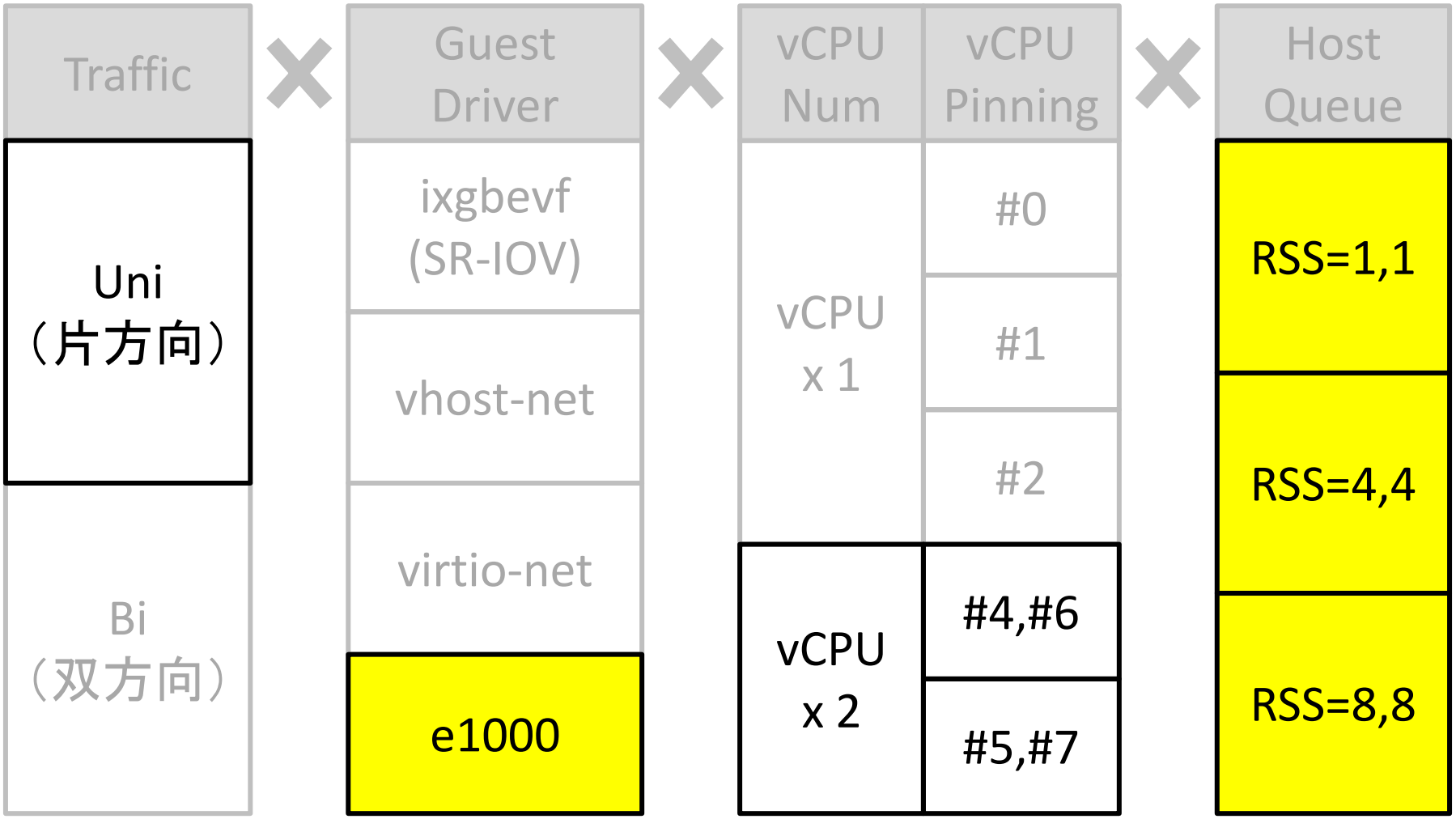
Host Queue（RSS）設定による性能差（Mbps）



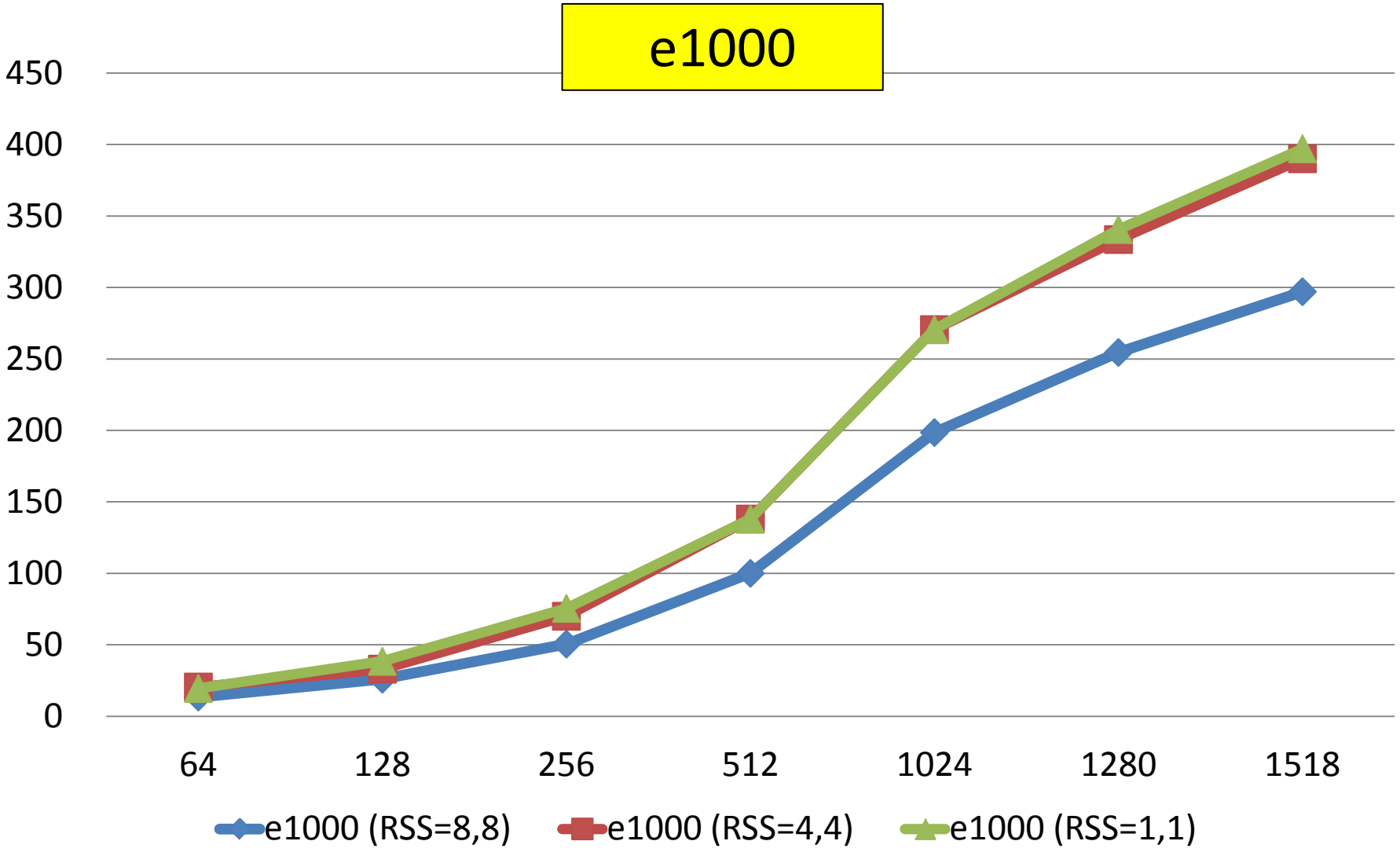
Host Queue（RSS）設定による性能差（fps）



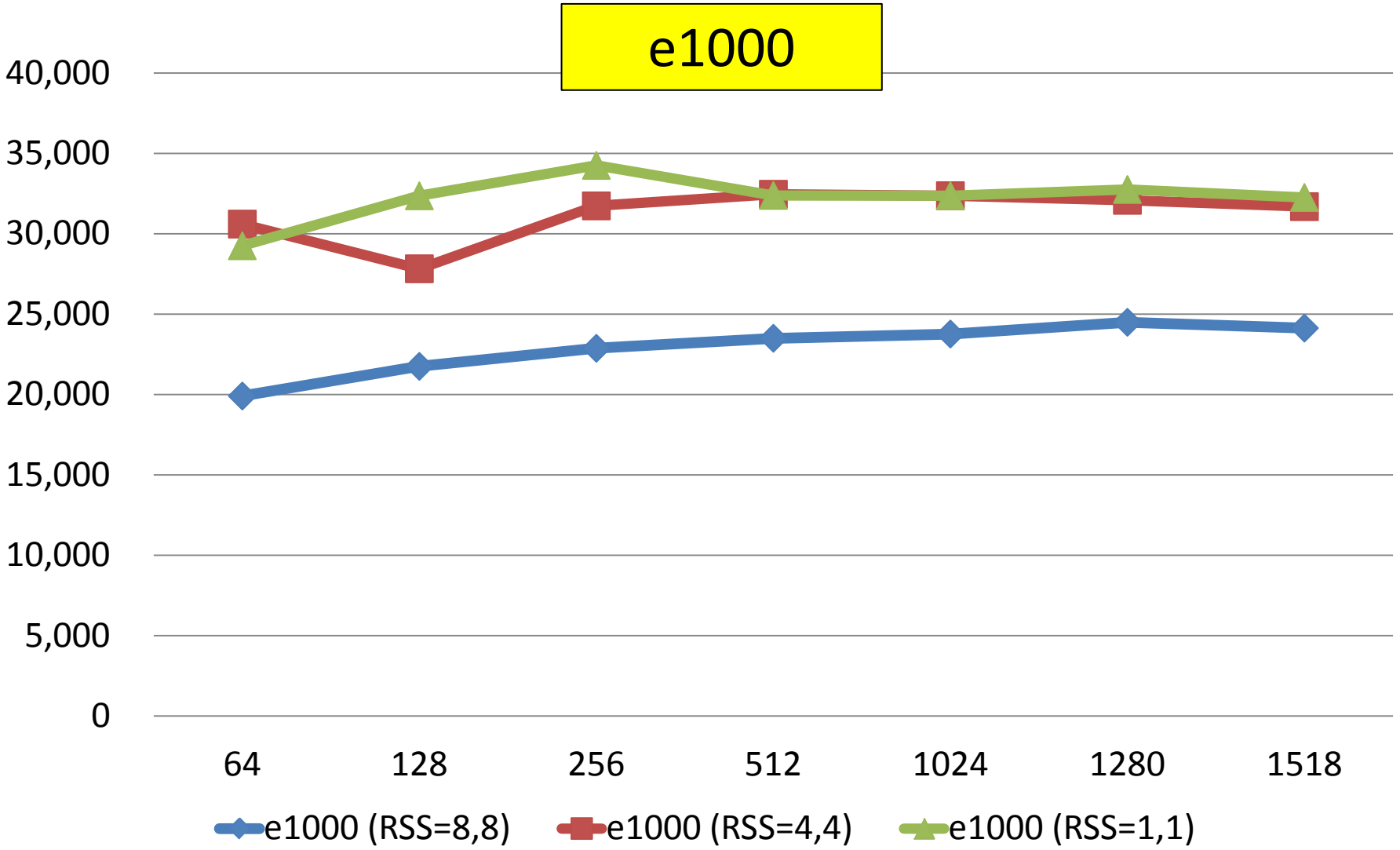
Host Queue（RSS）設定による性能差



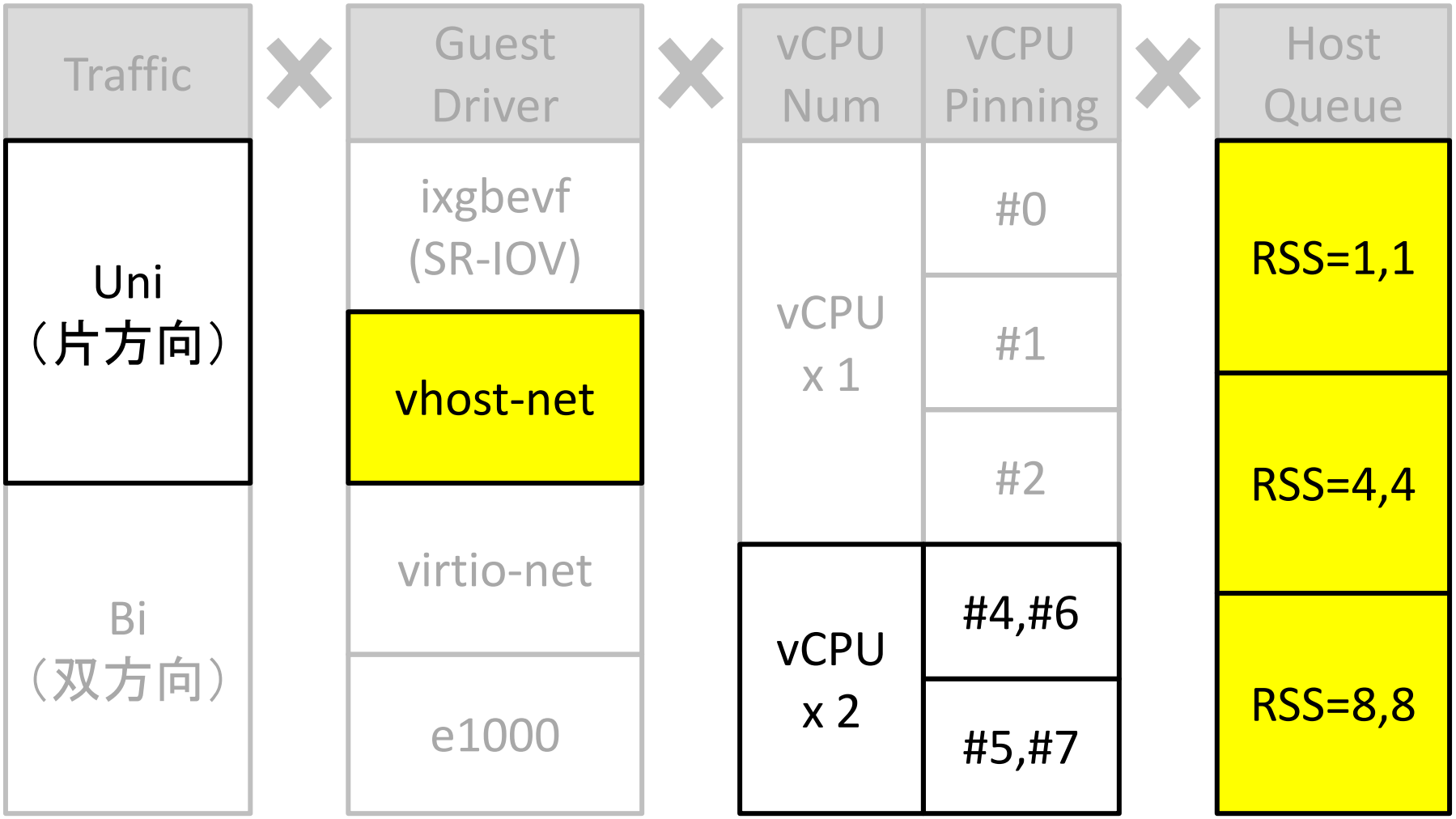
Host Queue（RSS）設定による性能差（Mbps）



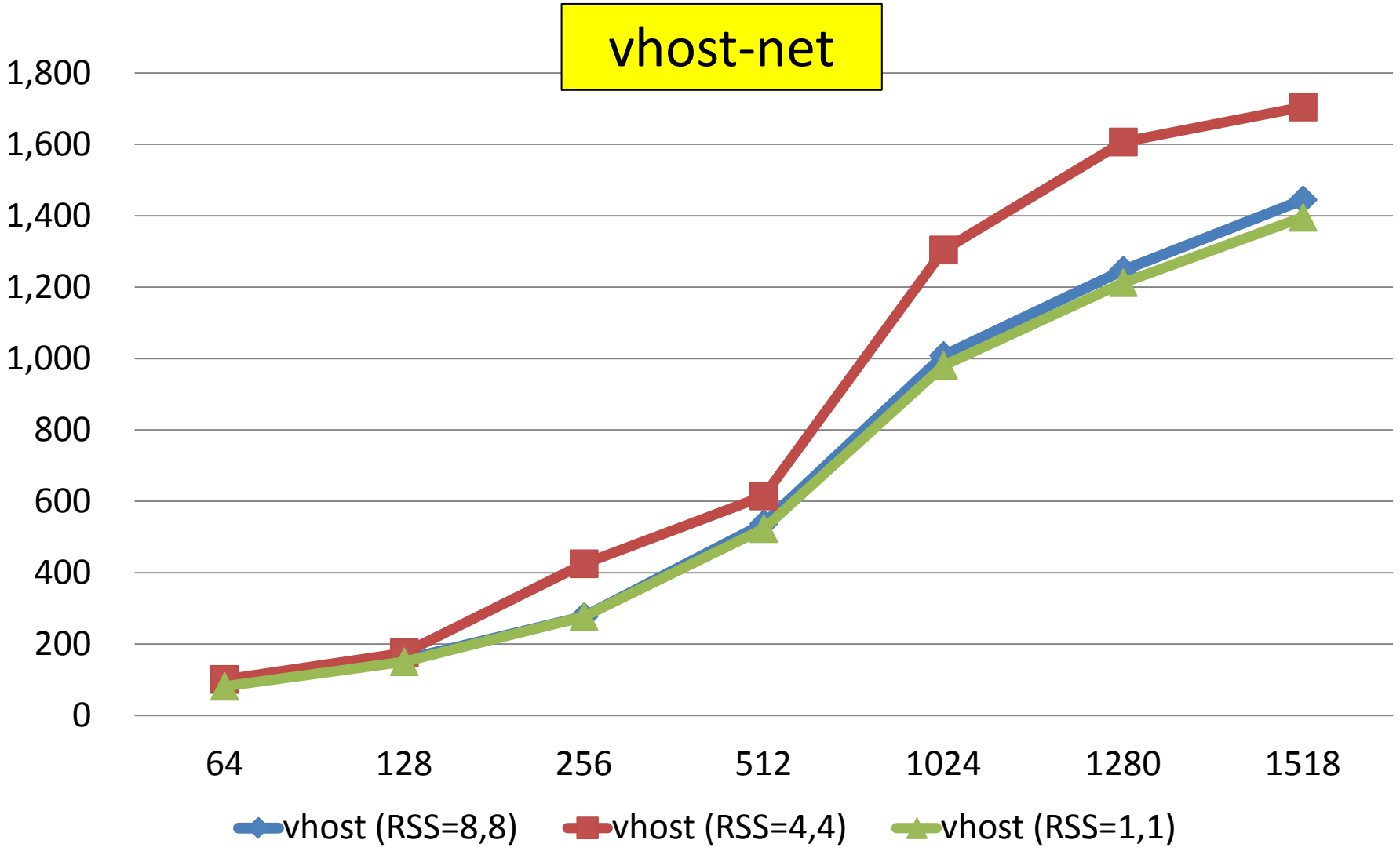
Host Queue（RSS）設定による性能差（fps）



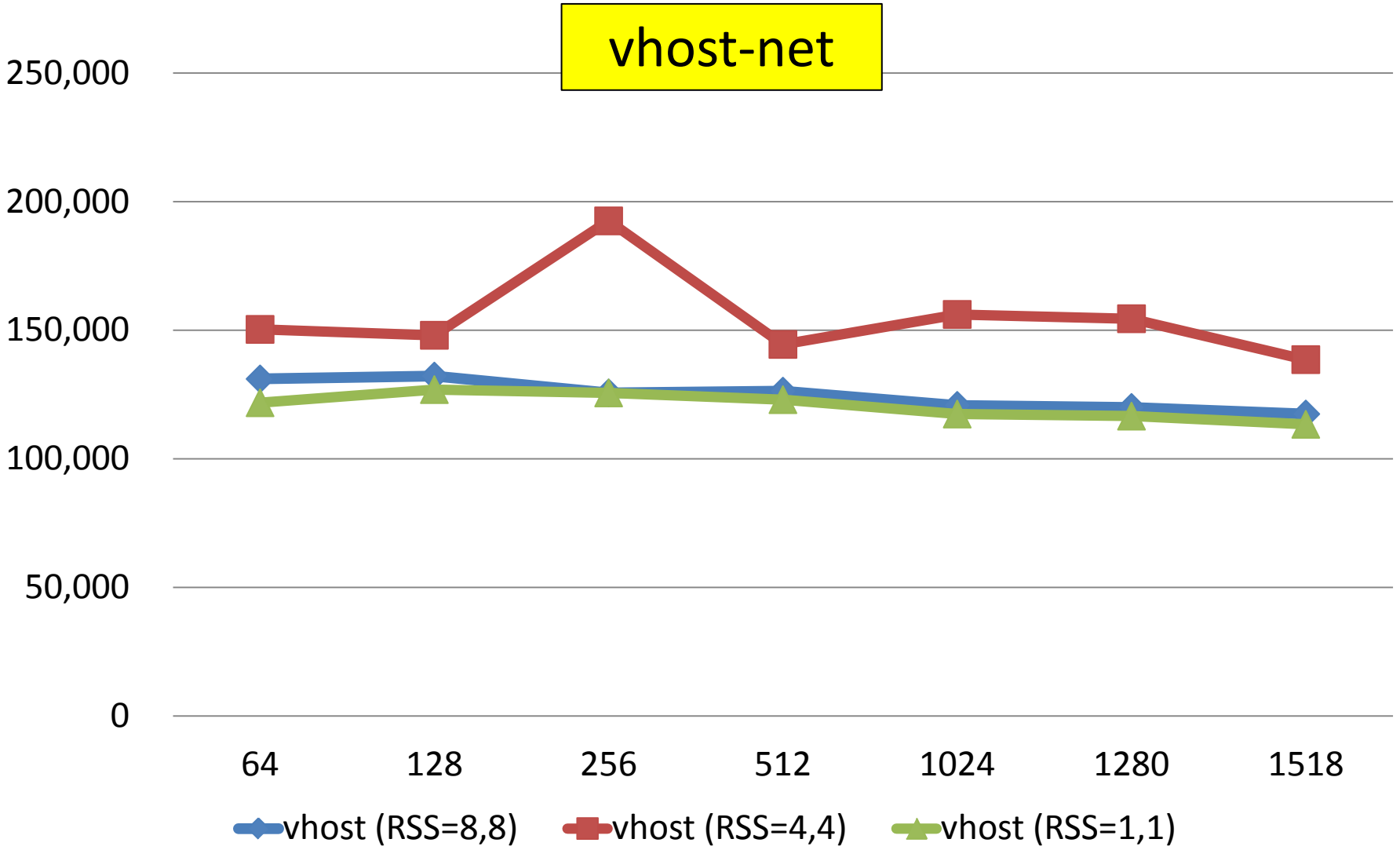
Host Queue（RSS）設定による性能差



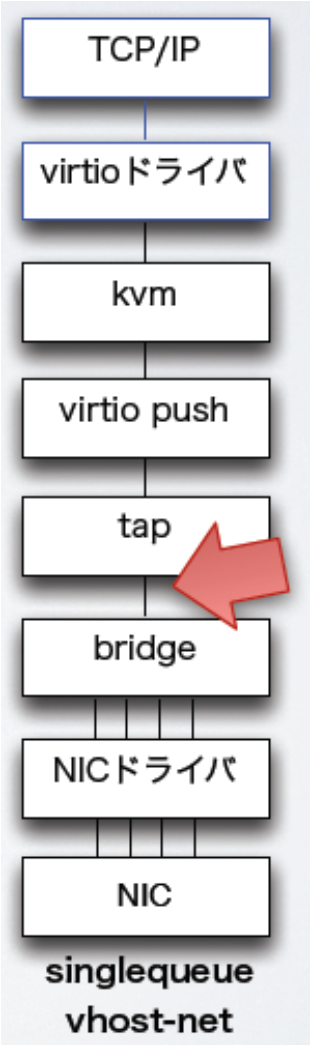
Host Queue（RSS）設定による性能差（Mbps）



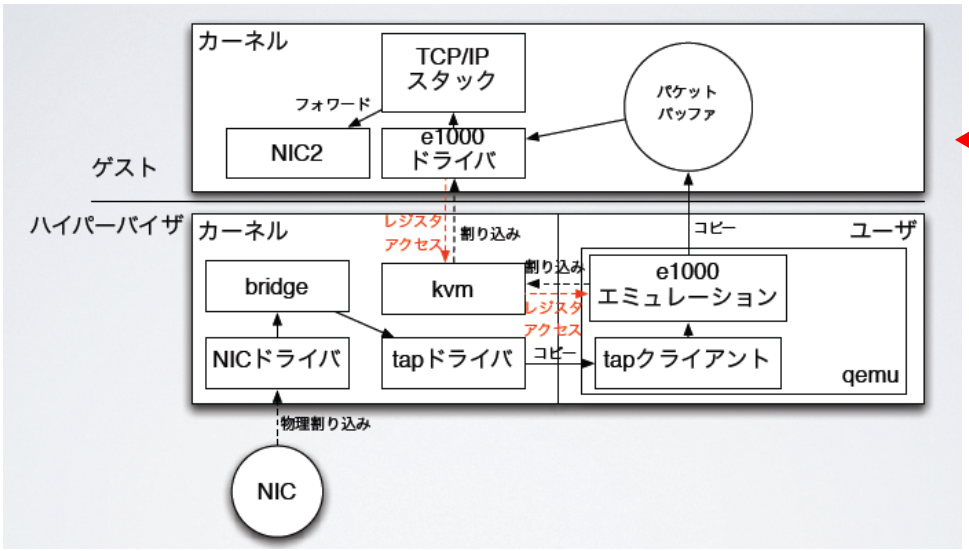
Host Queue（RSS）設定による性能差（fps）



Host Queue（RSS）設定による性能差（fps）



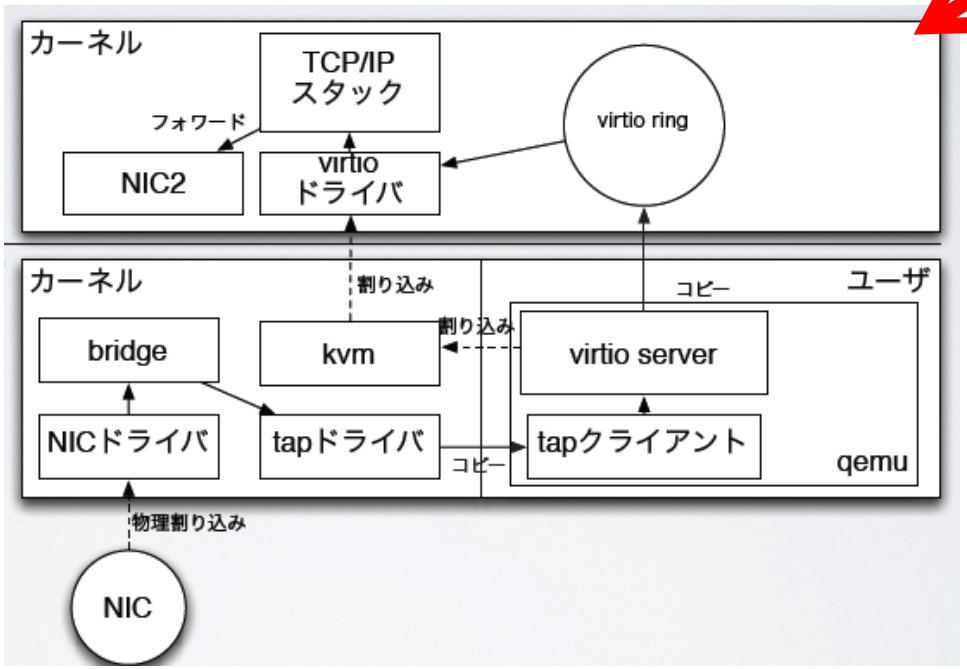
- virtio : RSS 1,1 >> 4,4 = 8,8
  - ロック競合による性能劣化？
  - qemu の main thread (通常) が eth0, eth1 処理
- e1000 : RSS 1,1 = 4,4 >> 8,8
  - ロック競合以外のオーバーヘッドが大きい？
- vhost: RSS 4,4 > 8,8 = 1,1
  - kernel thread でeth0 と eth1 で 2 本起動
  - 処理分散のメリットの方が大きい？？

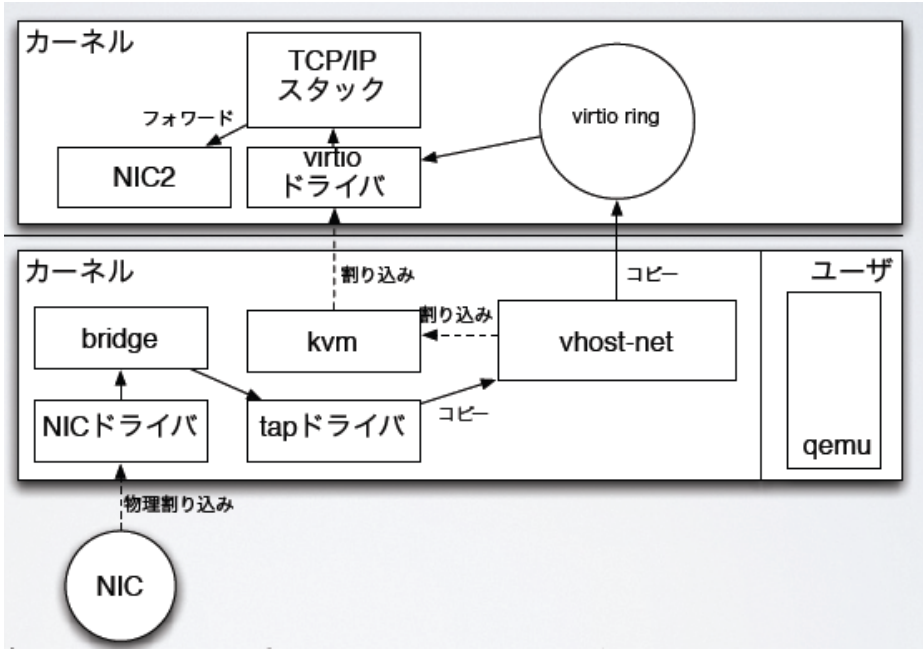


e1000

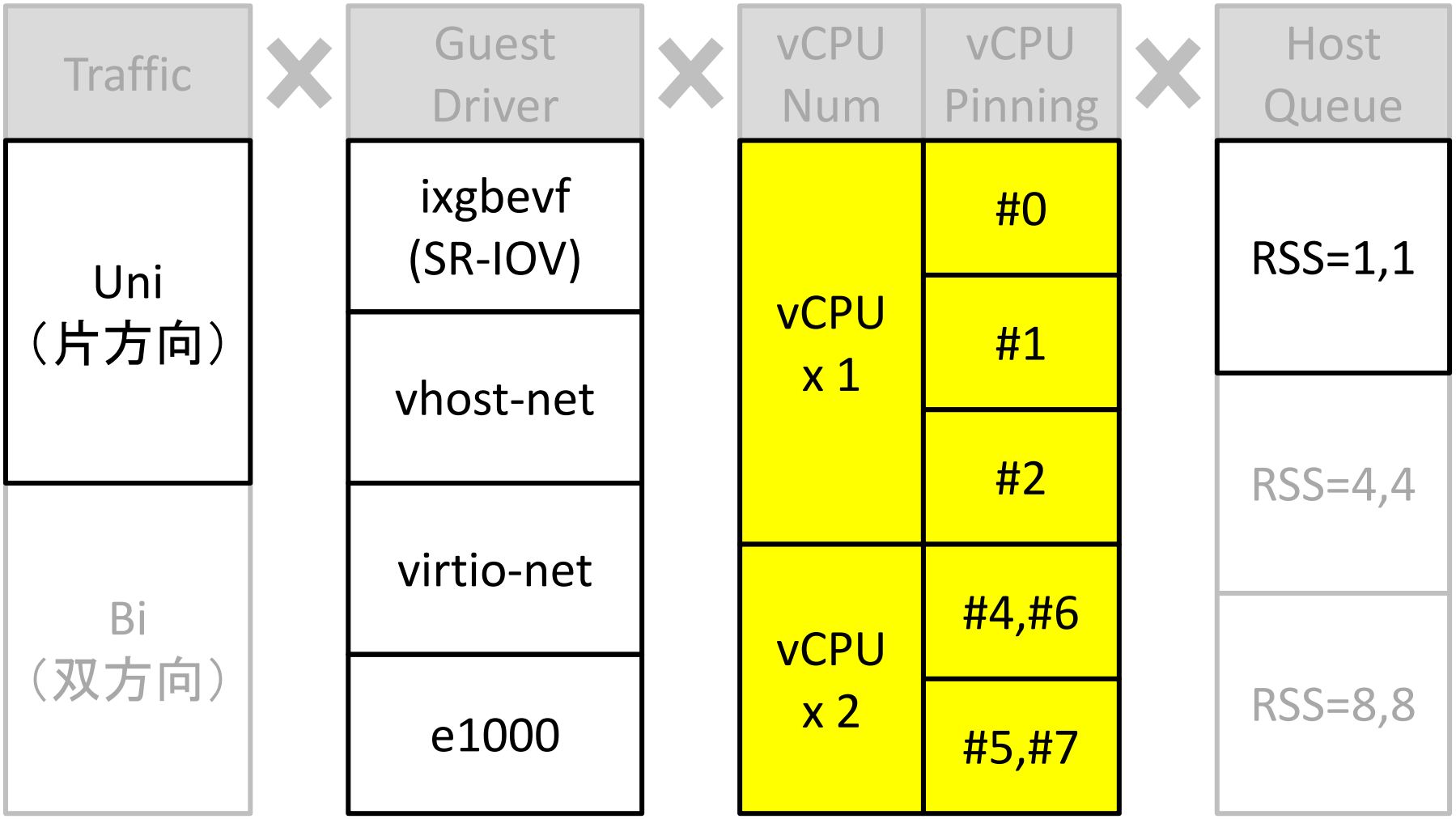
virtio

vhost

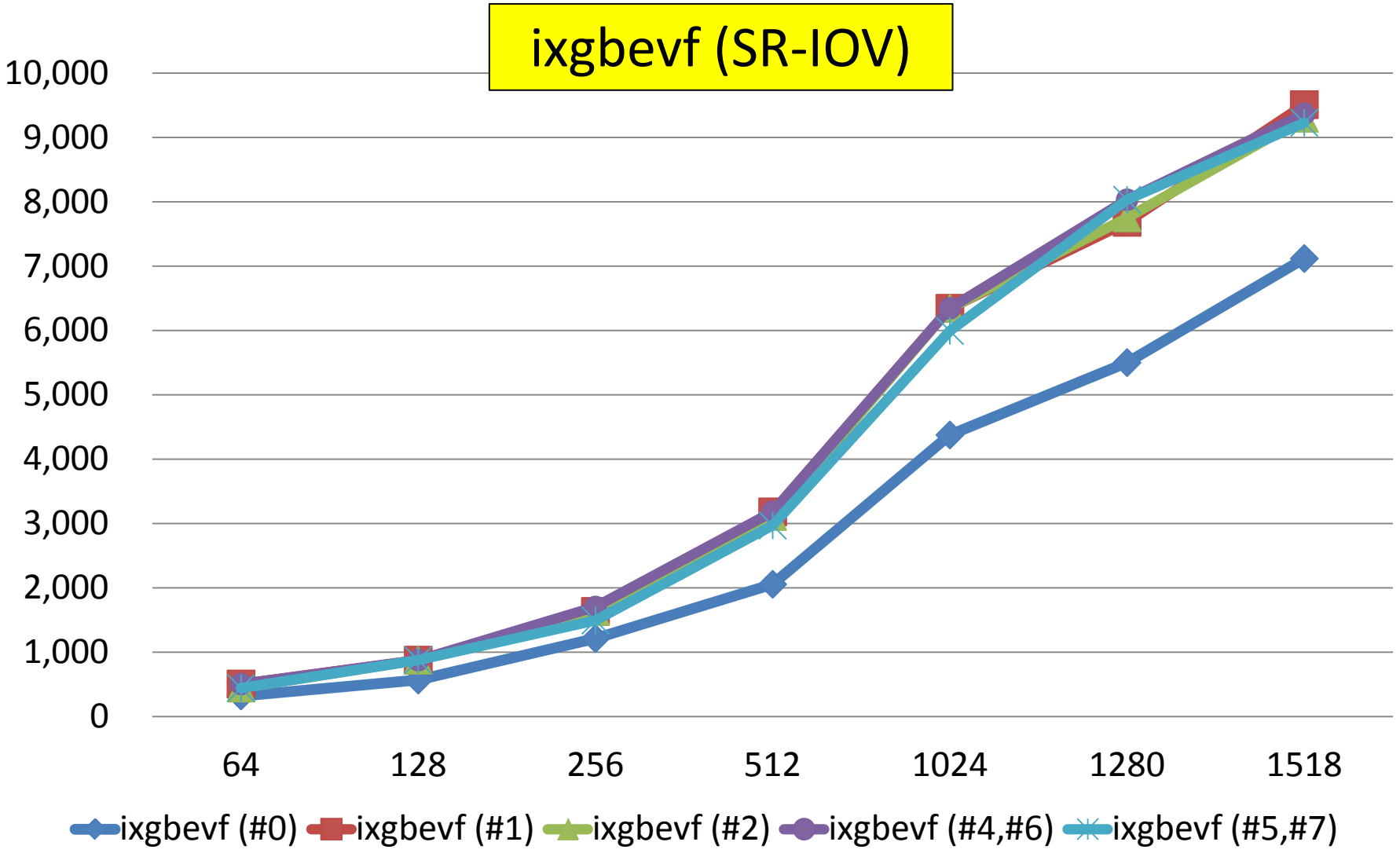




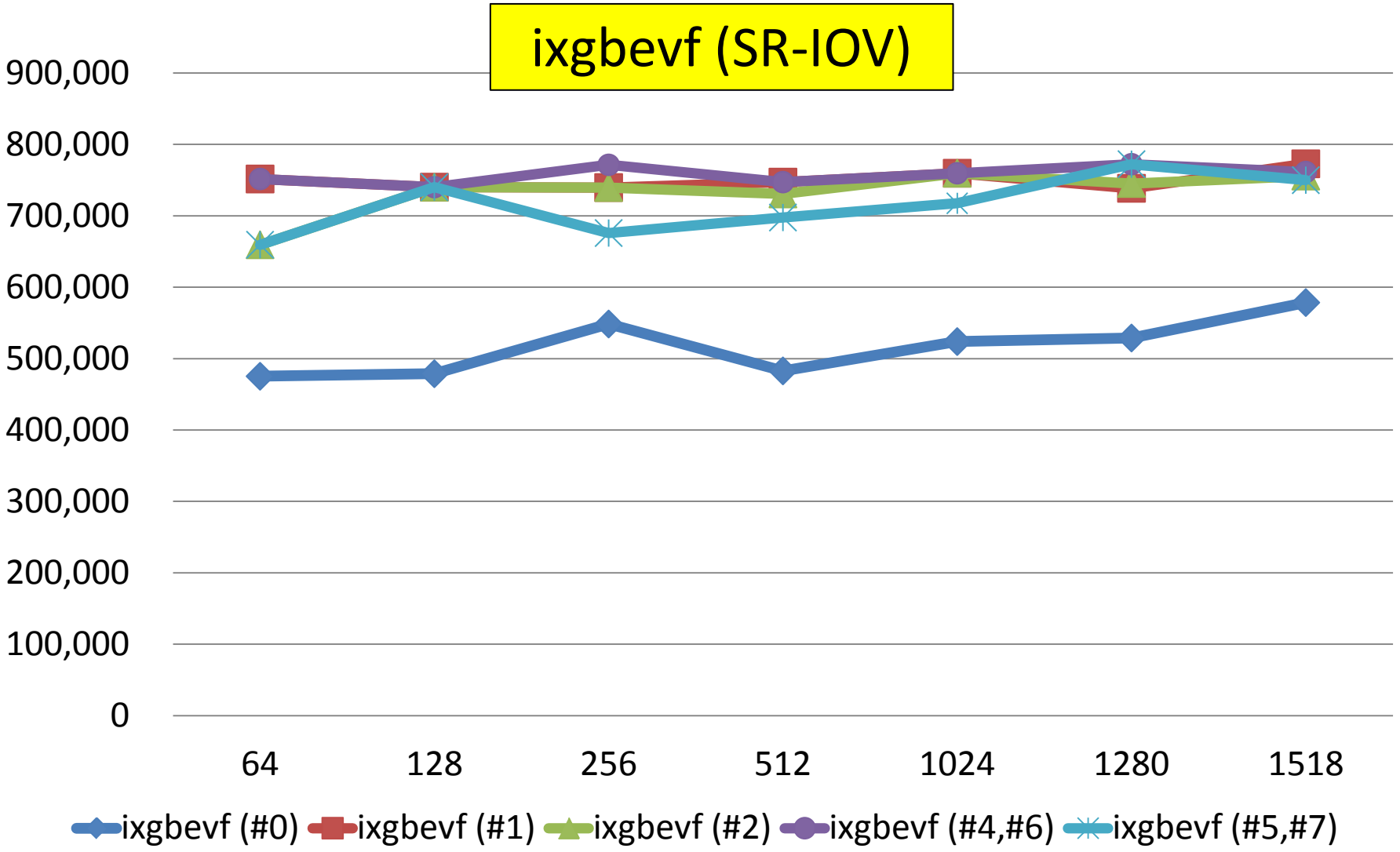
vCPU Num/Pinning による性能差



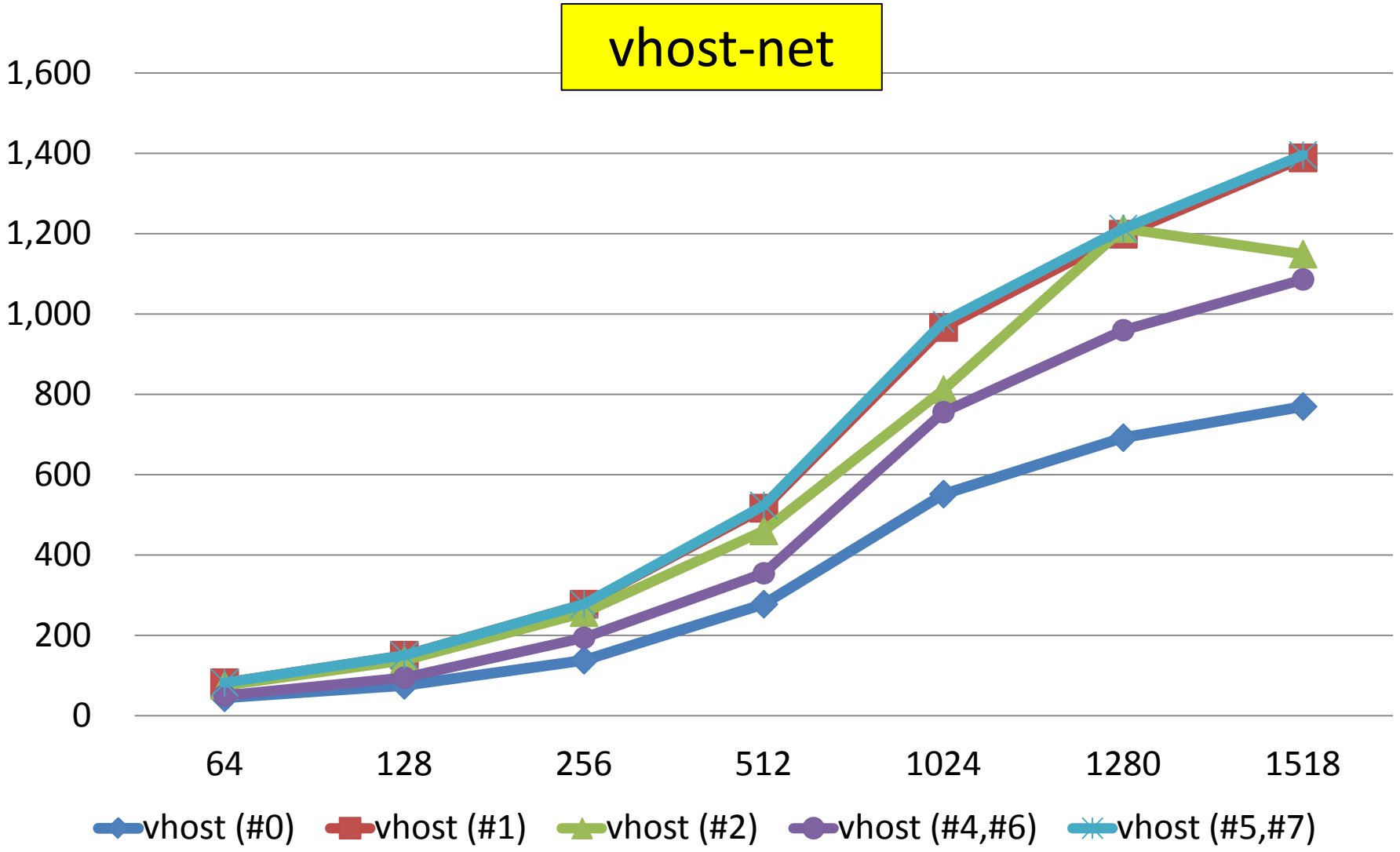
vCPU Num/Pinning による性能差（Mbps）



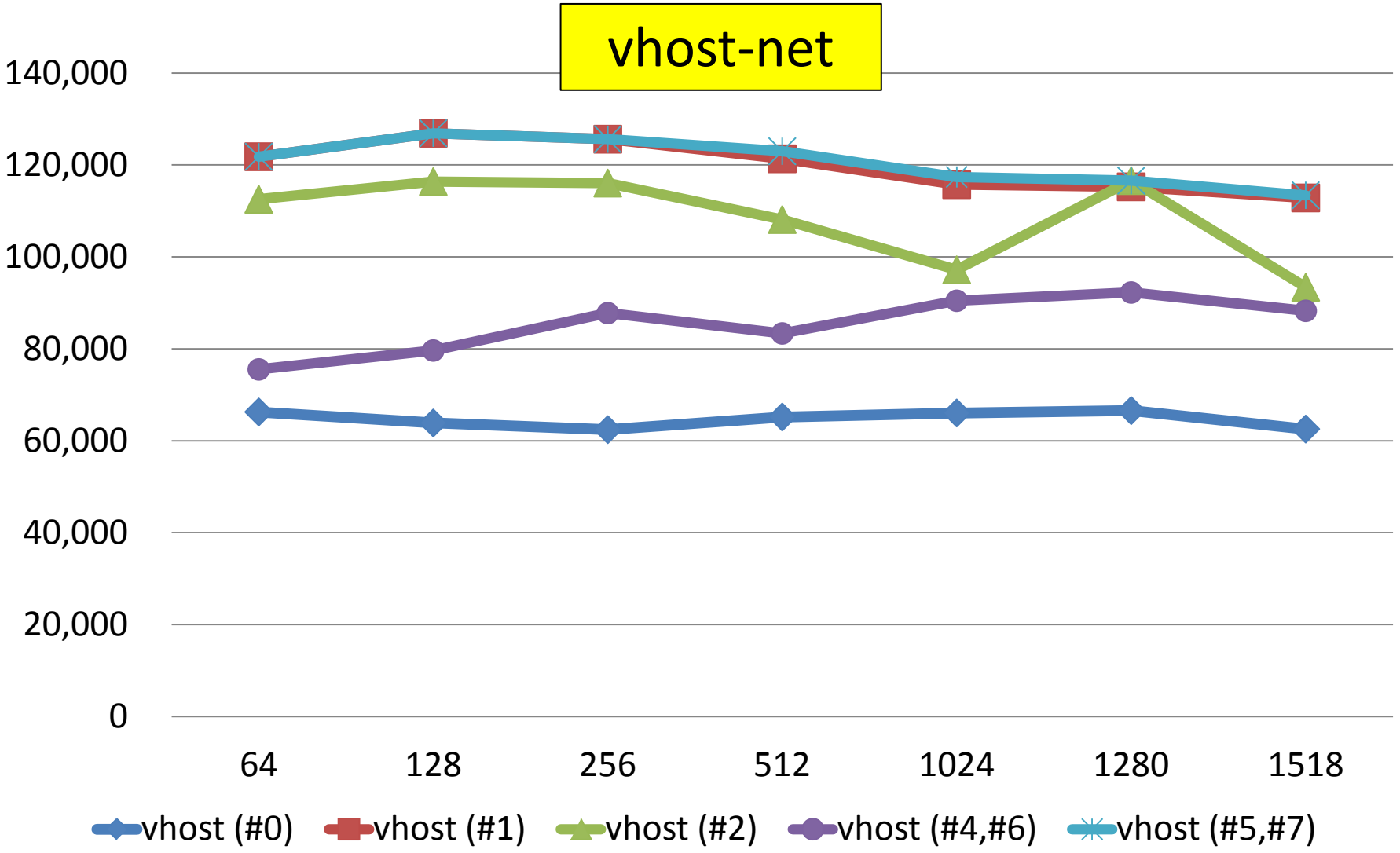
vCPU Num/Pinning による性能差（fps）



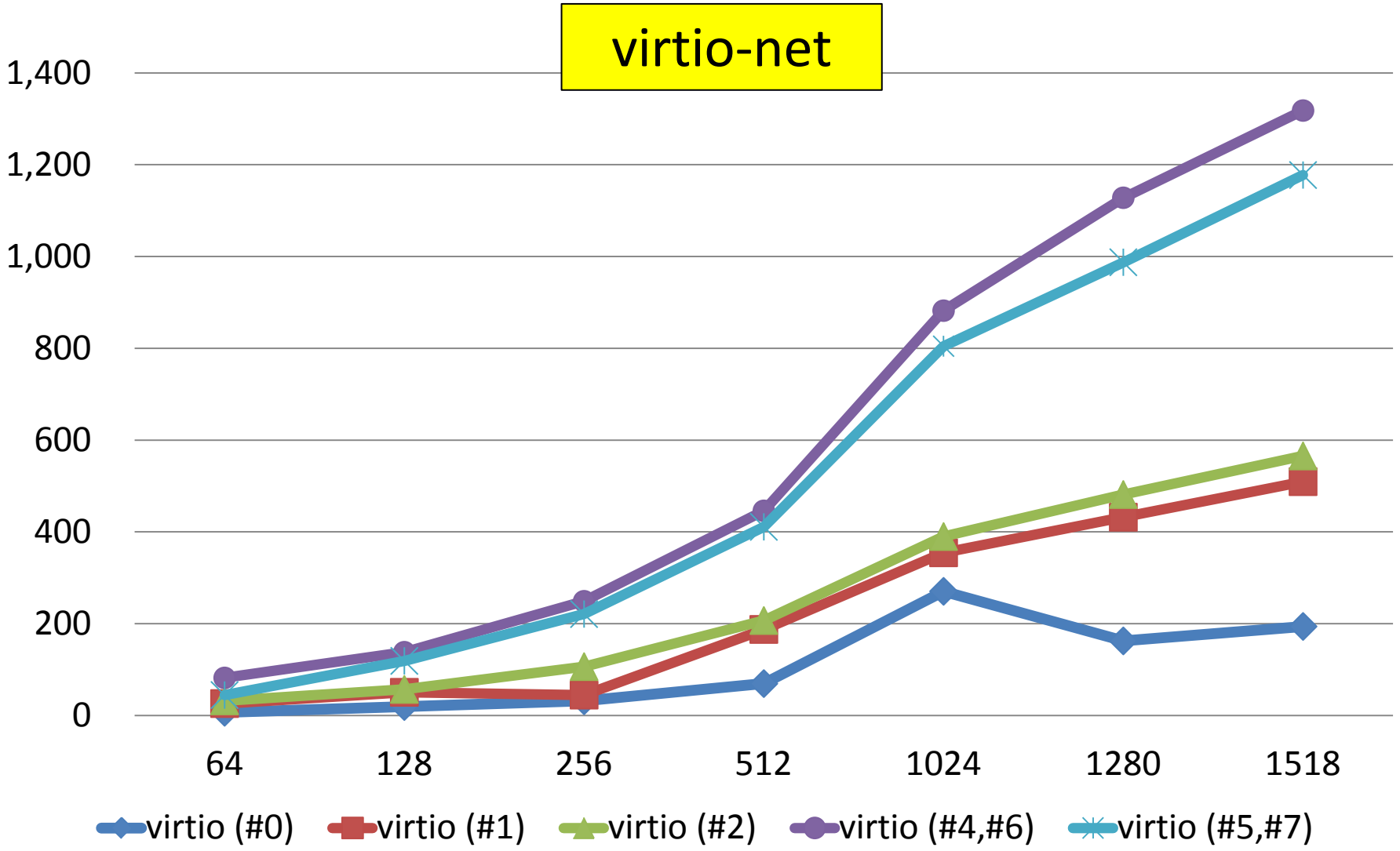
vCPU Num/Pinning による性能差（Mbps）



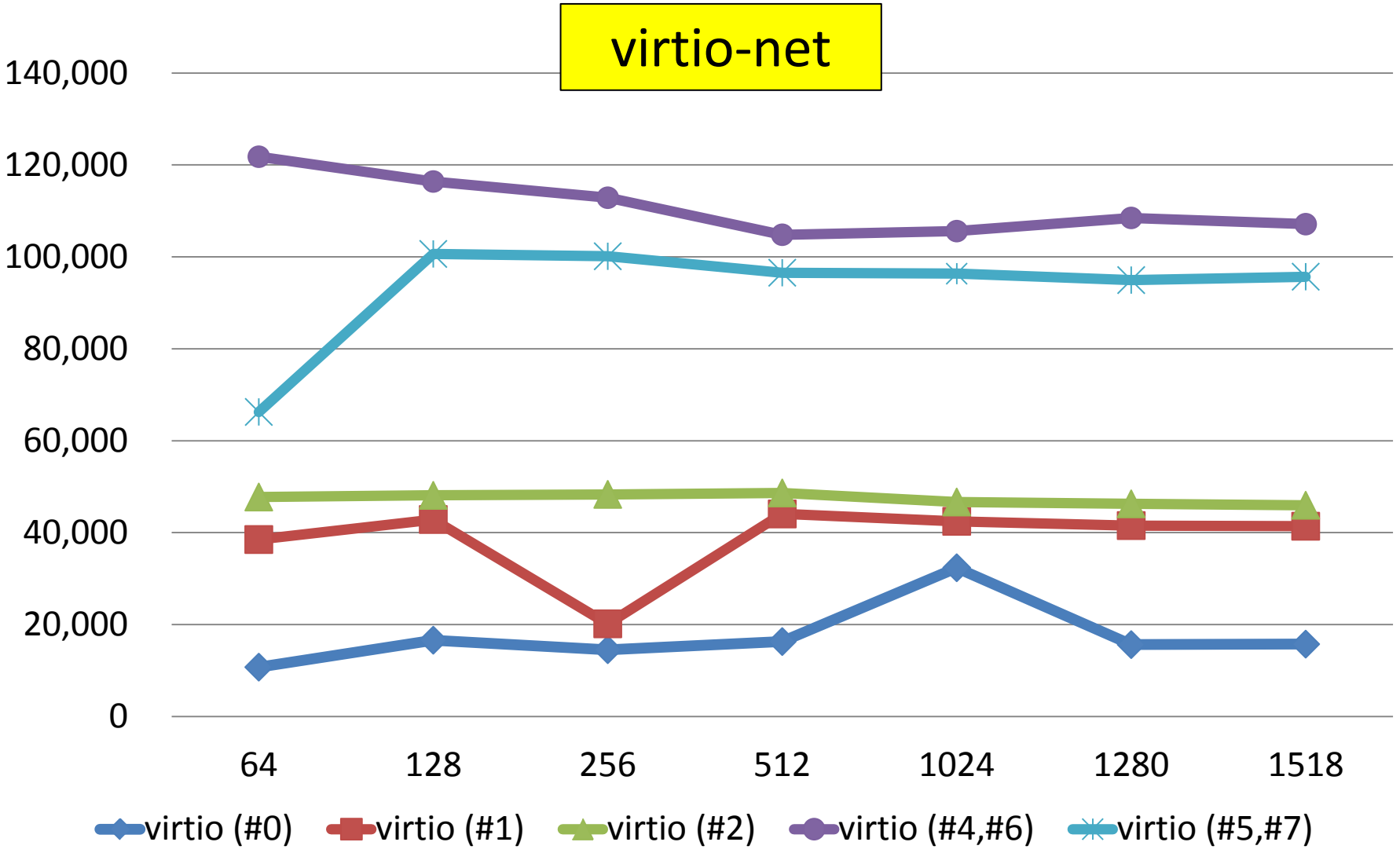
vCPU Num/Pinning による性能差（fps）



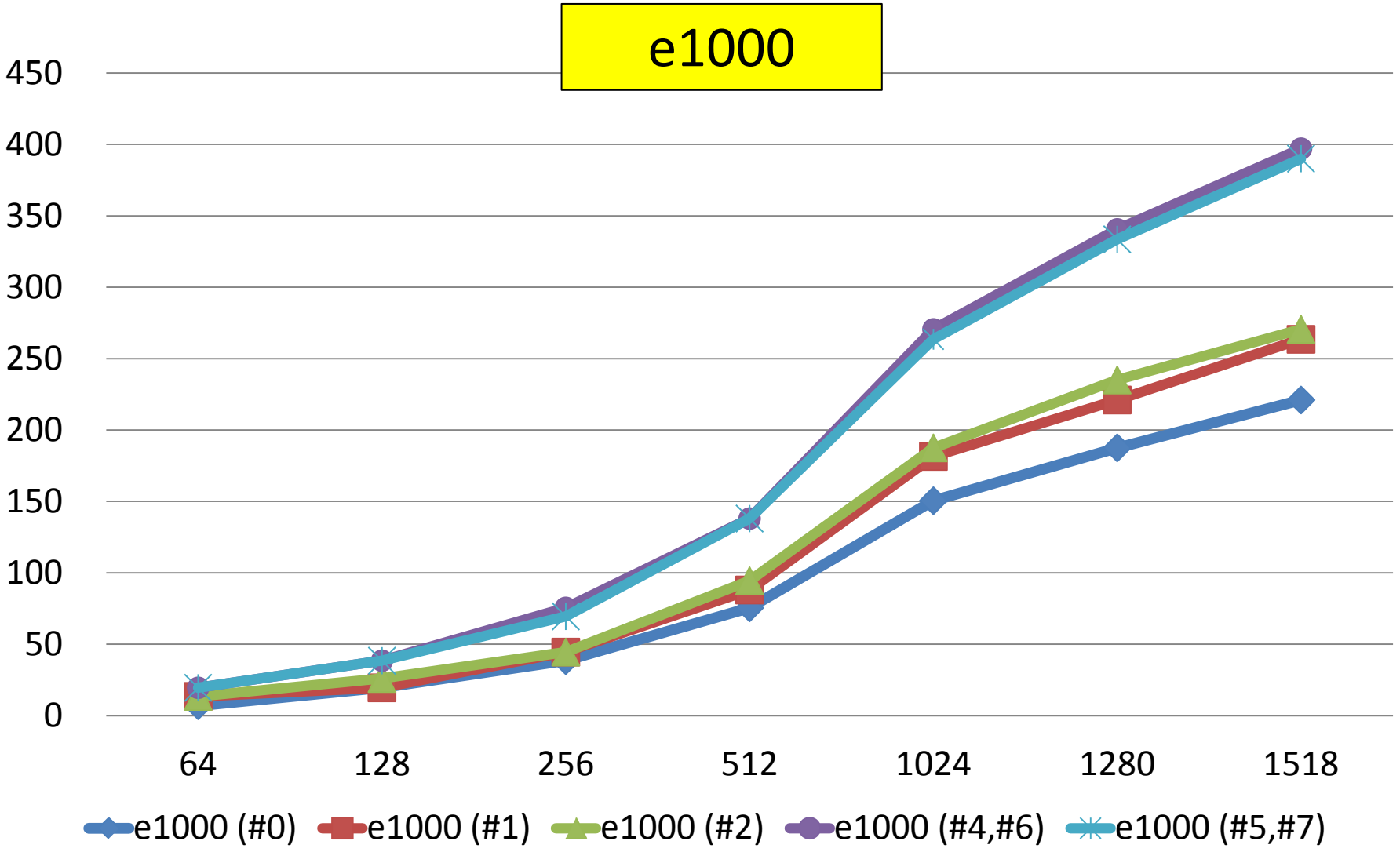
vCPU Num/Pinning による性能差（Mbps）



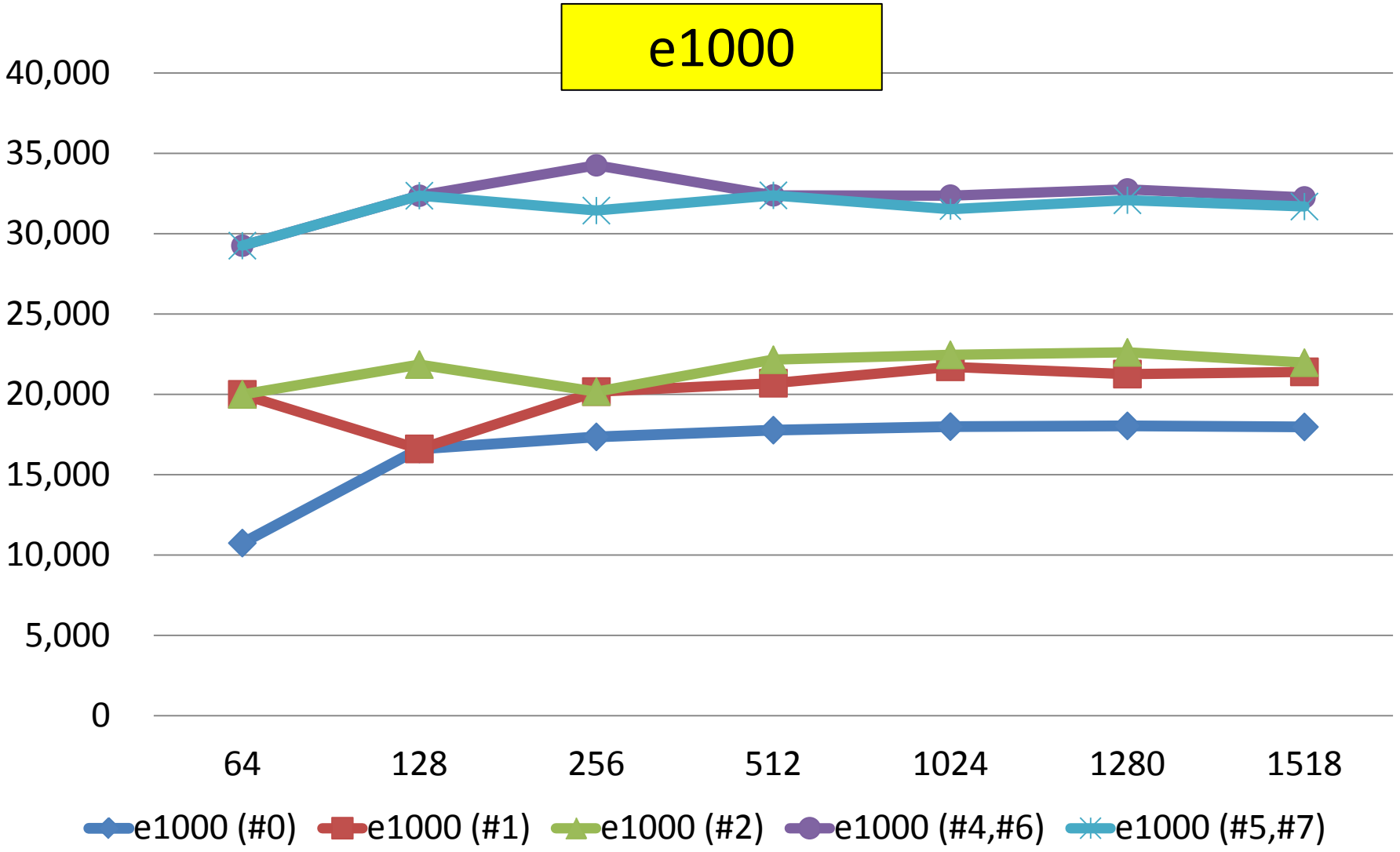
vCPU Num/Pinning による性能差（fps）



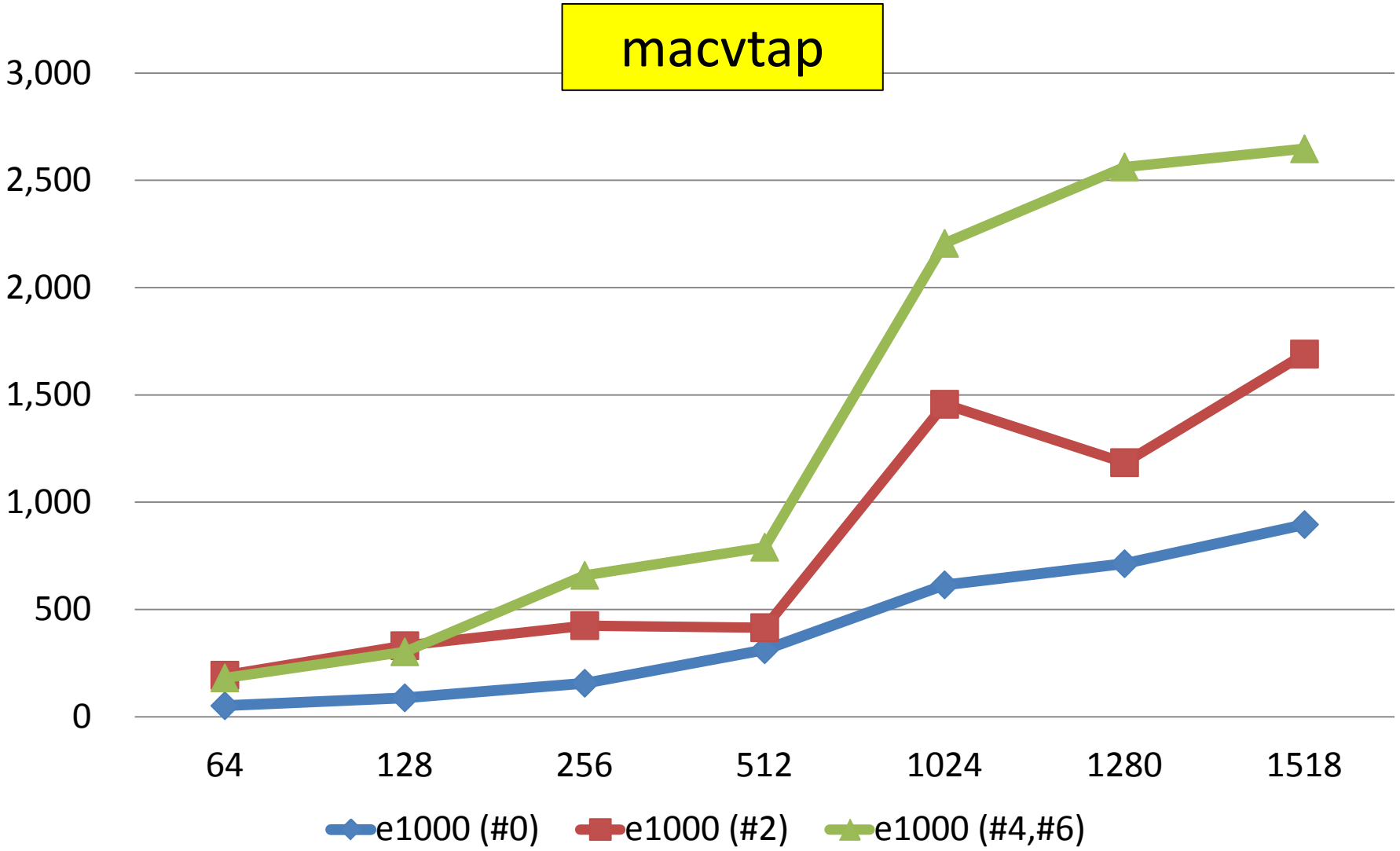
vCPU Num/Pinning による性能差（Mbps）



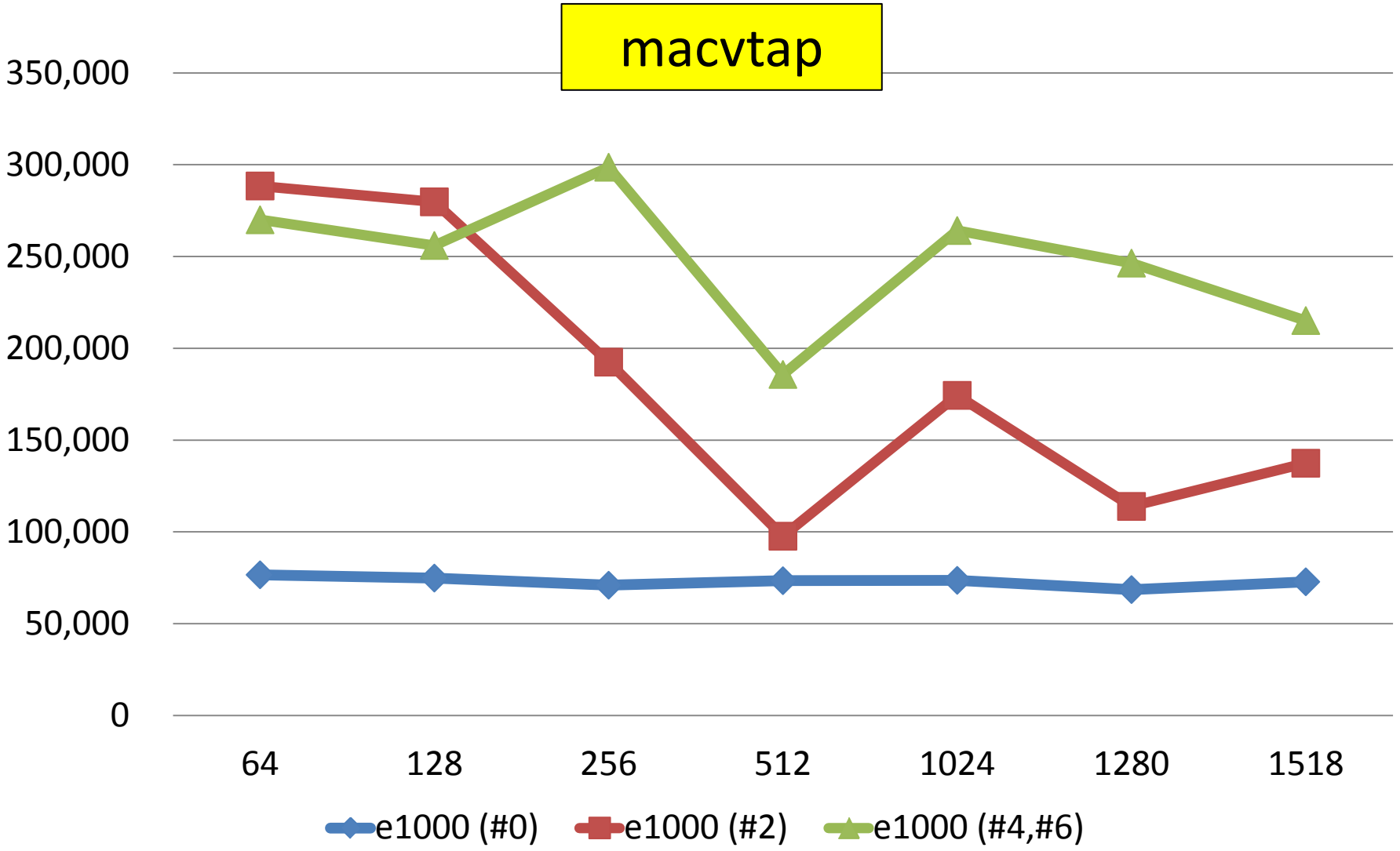
vCPU Num/Pinning による性能差（fps）



vCPU Num/Pinning による性能差（Mbps）

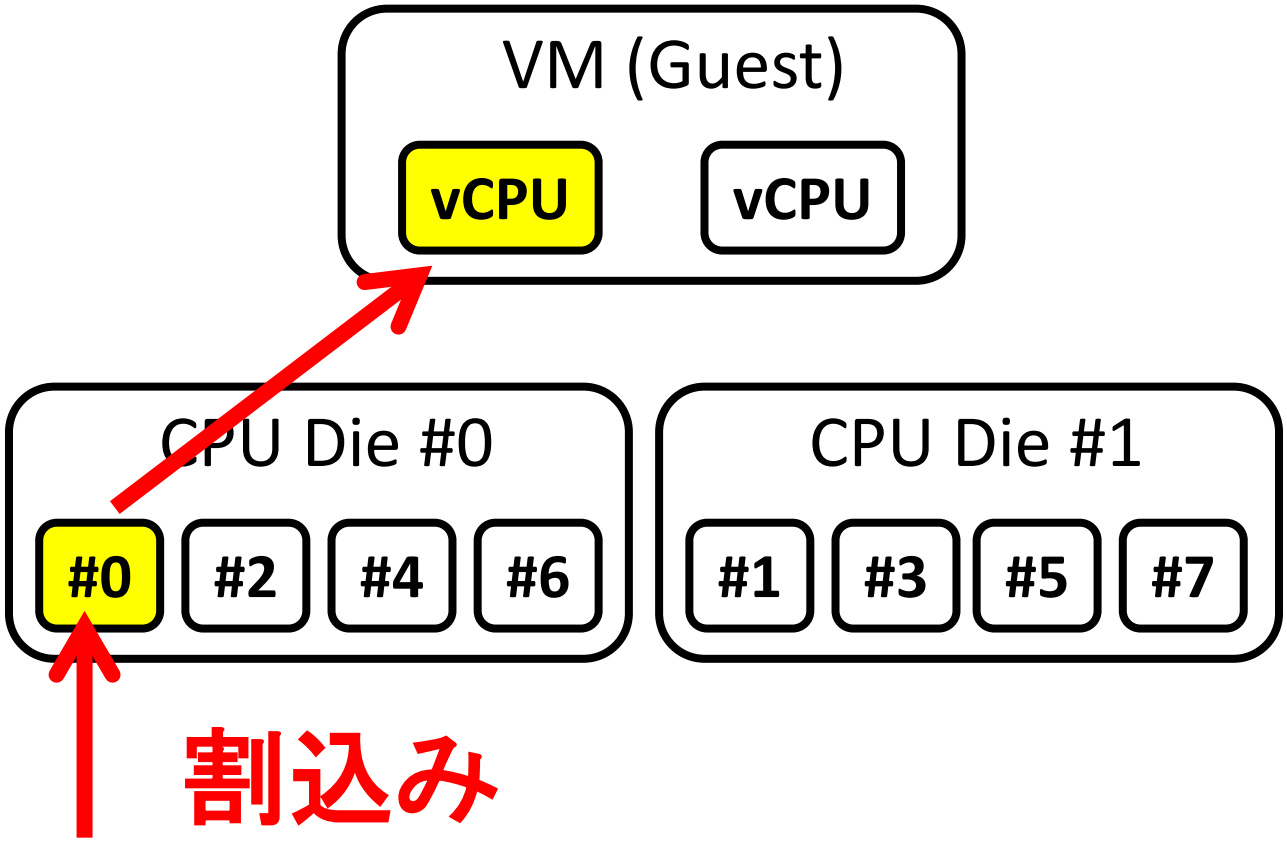


vCPU Num/Pinning による性能差（fps）



vCPU Num/Pinning による性能差（fps）

ホスト割込みCoreとVMのCoreが同じ場合、Core性能がボトルネックに

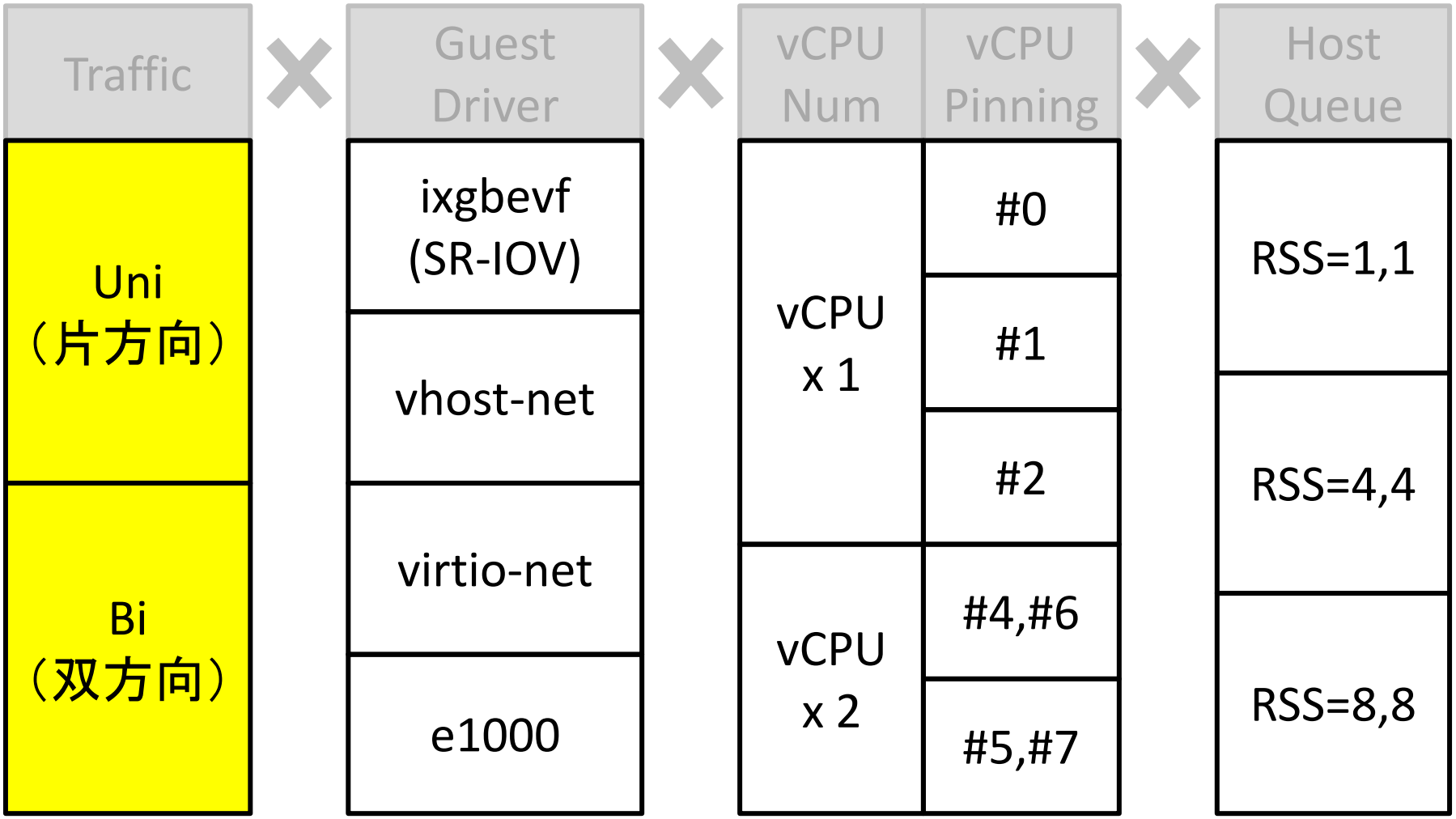


vCPU Num/Pinning による性能差（fps）

- vCPU数による性能は（pinning位置の影響）は Guest Driver により異なる
  - CPU0へ Pinning した場合を除く（前Slide）

- ixgbevf (SR-IOV)
  - Core増加で変わらず / pinning位置による影響小
- vhost-net
  - Core増加で変わらず / pinning位置による影響大
- virtio-net,e1000, macvtap
  - Core増加で向上 / pinning位置による影響小

Traffic Direction による性能差



Directionによる性能差（SR-IOV）

vCPU Pin	Dir	Ave	Min	Max	Ave	Min	Max
#0	Uni	516623.4	475260	578445			
#0	Bi	396494.6	330447	424674	77%	70%	73%
#1	Uni	749841.6	738431	772742			
#1	Bi	713379.3	582217	748522	95%	79%	97%
#2	Uni	732844.4	659412	759549			
#2	Bi	664425.1	582217	694070	91%	88%	91%
#4,#6	Uni	757587.3	740076	772236			
#4,#6	Bi	680312.4	582217	722580	90%	79%	94%
#5,#7	Uni	716059.6	659412	772236			
#5,#7	Bi	569646.4	513474	611478	80%	78%	79%

Directionによる性能差（vhost：RSS=8,8）

vCPU Pin	Direction	Ave	Min	Max	Ave	Min	Max
#0	Uni	109,045	73,681	121,791	129%	173%	127%
#0	Bi	140,815	127,745	154,802			
#1	Uni	88,234	55,063	119,240	160%	232%	131%
#1	Bi	140,848	127,745	155,945			
#2	Uni	87,029	66,267	122,424	159%	193%	125%
#2	Bi	137,984	127,745	153,395			
#4,#6	Uni	94,494	38,505	131,978	134%	152%	115%
#4,#6	Bi	126,632	58,500	152,043			
#5,#7	Uni	124,747	117,404	132,153	105%	97%	108%
#5,#7	Bi	130,516	114,025	143,373			

Directionによる性能差（vhost：RSS=4,4）

vCPU Pin	Direction	Ave	Min	Max	Ave	Min	Max
#0	Uni	176,099	162,549	192,376	99%	105%	92%
#0	Bi	174,416	170,608	176,589			
#2	Uni	130,100	122,424	141,090	123%	125%	119%
#2	Bi	160,132	153,274	168,517			
#4,#6	Uni	154,910	138,548	192,482	97%	102%	82%
#4,#6	Bi	150,940	141,087	157,335			

Directionによる性能差（vhost：RSS=1,1）

vCPU Pin	Direction	Ave	Min	Max	Ave	Min	Max
#0	Uni	64,694	62,415	66,557	123%	121%	124%
#0	Bi	79,381	75,221	82,798			
#1	Uni	119,916	112,832	126,901	80%	83%	77%
#1	Bi	95,552	93,900	97,289			
#2	Uni	108,612	93,402	116,587	88%	101%	83%
#2	Bi	95,552	93,900	97,289			
#4,#6	Uni	85,316	75,521	92,248	112%	124%	105%
#4,#6	Bi	95,552	93,900	97,289			
#5,#7	Uni	120,667	113,403	126,901	79%	82%	77%
#5,#7	Bi	95,389	93,085	97,289			

Directionによる性能差（virtio：RSS=8,8）

vCPU Pin	Direction	Ave	Min	Max	Ave	Min	Max
CPU0	Uni	38,520	37,272	40,017	116%	107%	117%
CPU0	Bi	44,518	39,807	46,992			
CPU1	Uni	37,557	32,200	40,017	119%	124%	117%
CPU1	Bi	44,518	39,807	46,992			
CPU2	Uni	38,780	37,426	40,017	115%	106%	117%
CPU2	Bi	44,518	39,807	46,992			
CPU4,6	Uni	48,236	37,426	75,149	97%	106%	72%
CPU4,6	Bi	46,900	39,807	53,948			
CPU5,7	Uni	48,199	35,307	75,149	93%	64%	71%
CPU5,7	Bi	44,816	22,593	53,601			

Directionによる性能差（virtio：RSS=1,1）

vCPU Pin	Direction	Ave	Min	Max	Ave	Min	Max
CPU0	Uni	17,421	10,742	32,365	142%	113%	124%
CPU0	Bi	24,678	12,172	39,992			
CPU1	Uni	38,716	20,168	44,070	130%	198%	123%
CPU1	Bi	50,224	39,992	54,212			
CPU2	Uni	47,396	45,971	48,645	128%	127%	133%
CPU2	Bi	60,751	58,500	64,717			
CPU4,6	Uni	111,010	104,815	121,791	56%	55%	53%
CPU4,6	Bi	62,177	57,235	64,762			
CPU5,7	Uni	92,940	66,267	100,639	64%	75%	64%
CPU5,7	Bi	59,058	49,654	64,717			

Directionによる性能差（e1000：RSS=8,8）

vCPU Pin	Direction	Ave	Min	Max	Ave	Min	Max
#0	Uni	20,018	16,523	21,234	115%	129%	113%
#0	Bi	23,055	21,391	23,946			
#1	Uni	21,054	19,904	22,043	110%	107%	110%
#1	Bi	23,218	21,391	24,255			
#2	Uni	20,322	16,523	22,043	113%	129%	109%
#2	Bi	23,055	21,391	23,946			
#4,#6	Uni	21,738	19,904	23,496	150%	107%	152%
#4,#6	Bi	32,510	21,391	35,732			
#5,#7	Uni	22,914	19,904	24,489	144%	107%	146%
#5,#7	Bi	33,029	21,391	35,732			

Directionによる性能差（e1000：RSS=1,1）

vCPU Pin	Direction	Ave	Min	Max	Ave	Min	Max
#0	Uni	16,637	10,742	18,035	129%	166%	127%
#0	Bi	21,544	17,804	22,915			
#1	Uni	20,261	16,602	21,713	125%	129%	128%
#1	Bi	25,227	21,484	27,698			
#2	Uni	21,601	19,996	22,611	127%	107%	133%
#2	Bi	27,334	21,484	30,023			
#4,#6	Uni	32,231	29,250	34,251	130%	137%	128%
#4,#6	Bi	41,810	39,992	43,708			
#5,#7	Uni	31,530	29,250	32,381	133%	137%	137%
#5,#7	Bi	41,871	39,992	44,305			

まとめ

- 目的にあった計測方法で
  ⇒ 限界性能 vs 機器・技術特性
- パケット転送性能には Guest Driver 選択が大きく影響
  ixgbevf(SR-IOV) >>> macvtap >>
  vhost-net > virtio-net >> e1000
- ホスト割込みと、VM割り当ては別コアに…
- コア数やホストキュー数（RSS）は、増やせばいいってもんじゃないよ

今後の課題

- 実運用に沿った調査を深堀り
  - CPU Pinningする？Auto Balanceにした時の性能は？
- VM数増加による転送性能の変化とGuest Driver毎の特性
- Bidirectional の時の評価
  - 傾向が Guest Driverにより異なる理由？？
- その他パラメーター変化による調査
  - リングサイズ、ゲストVM数