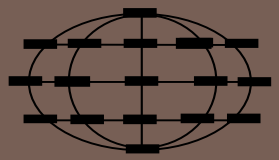
Институт Вычислительной Математики и Математической Геофизики
лаборатория Синтеза Параллельных Программы



О влиянии системных прерываний на производительность параллельных программ

Константин Калгин
kalgin@ssd.sscc.ru

Аннотация

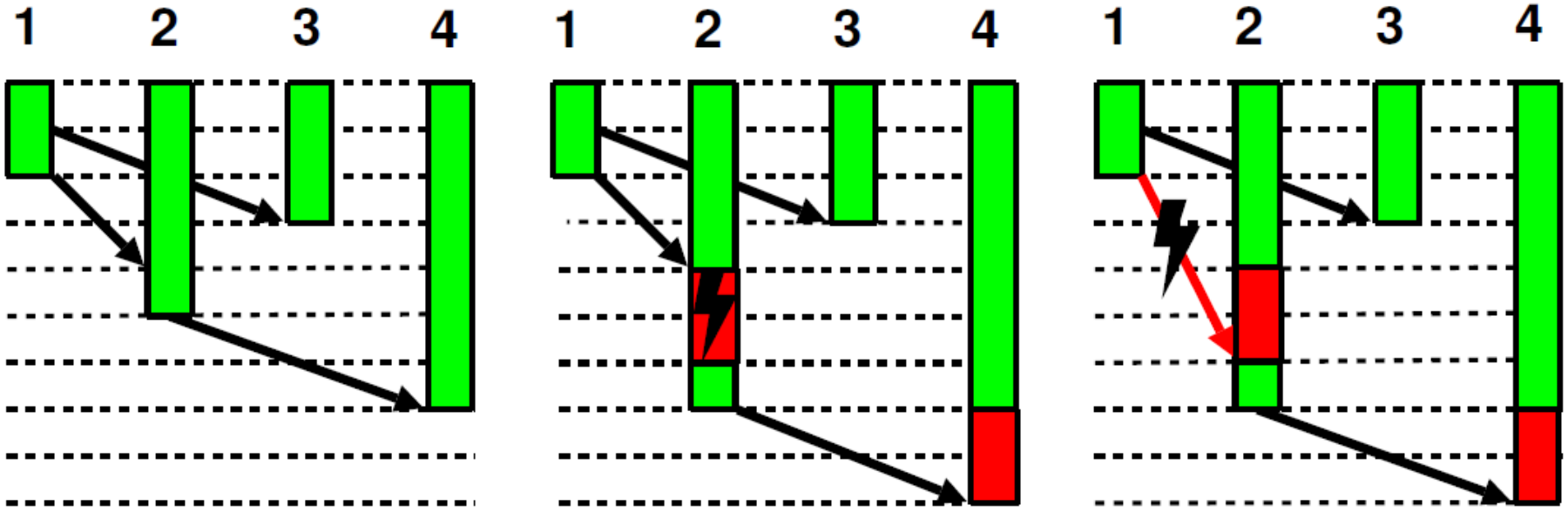
В работе исследуется влияние наиболее затратных по ресурсам регулярных системных прерываний (таймер и планировщик).

Эти прерывания, в зависимости от аппаратной архитектуры и настроек операционной системы, занимают 0.1-5% времени работы CPU, но могут стать причиной 10-100% ухудшения производительности параллельной программы (например, в массовых операциях типа MPI_Reduce).

Исследуется влияние этих прерываний на время работы класса параллельных программ с синхронизацией между «соседними» процессами на каждой итерации (например, трафаретные вычисления, синхронный клеточный автомат, явная разностная схема).

Строится модель распространения "системного шума" в параллельно работающих процессах. Приводятся результаты тестирования на вычислительных кластерах. Формулируются меры по минимизации влияния прерываний на производительность параллельной программы.

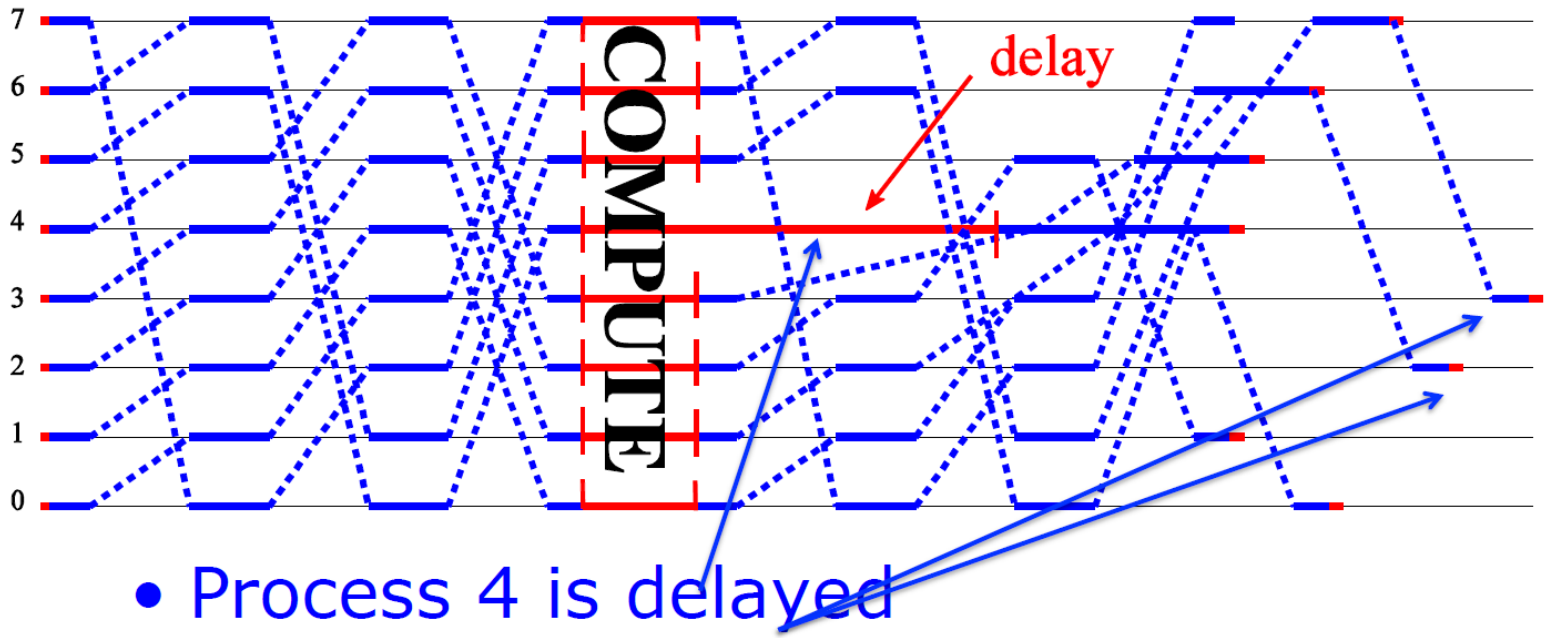
Noise : OS / Net



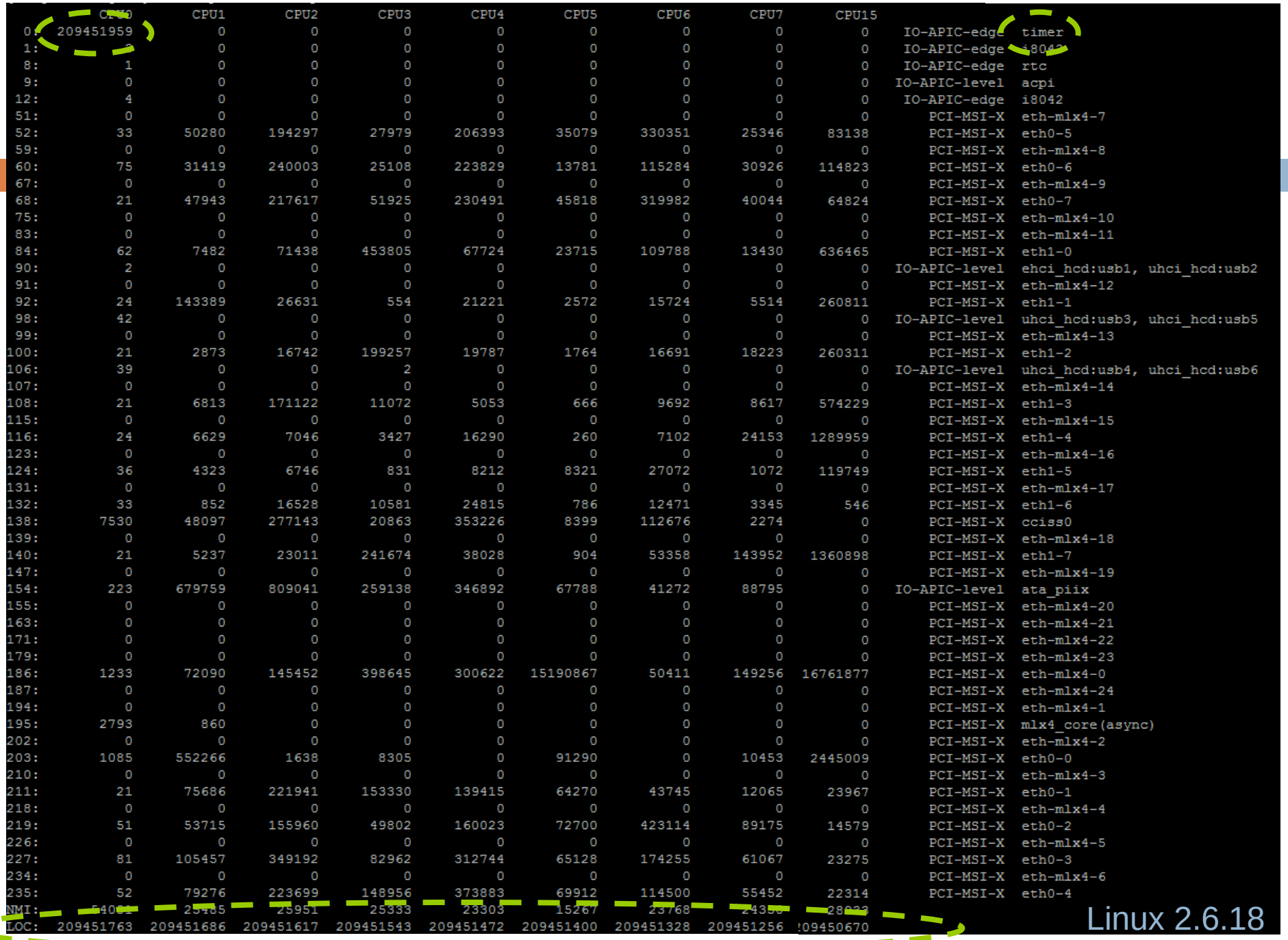
(a) No Noise (b) OS Noise (c) Net Noise

Барьер

A Noisy Example – Dissemination Barrier



- Process 4 is delayed
  - Noise propagates "*wildly*" (of course deterministic)


CPU0 CPU1 CPU2 CPU3 CPU4 CPU5 CPU6 CPU7 CPU15
0: 209451959 0 0 0 0 0 0 0 0 IO-APIC-edge timer
1: 0 0 0 0 0 0 0 0 0 IO-APIC-edge i8042
8: 1 0 0 0 0 0 0 0 0 IO-APIC-edge rtc
9: 0 0 0 0 0 0 0 0 0 IO-APIC-level acpi
12: 4 0 0 0 0 0 0 0 0 IO-APIC-edge i8042
51: 0 0 0 0 0 0 0 0 0 PCI-MSI-X eth-mlx4-7
52: 33 50280 194297 27979 206393 35079 330351 25346 83138 PCI-MSI-X eth0-5
59: 0 0 0 0 0 0 0 0 0 PCI-MSI-X eth-mlx4-8
60: 75 31419 240003 25108 223829 13781 115284 30926 114823 PCI-MSI-X eth0-6
67: 0 0 0 0 0 0 0 0 0 PCI-MSI-X eth-mlx4-9
68: 21 47943 217617 51925 230491 45818 319982 40044 64824 PCI-MSI-X eth0-7
75: 0 0 0 0 0 0 0 0 0 PCI-MSI-X eth-mlx4-10
83: 0 0 0 0 0 0 0 0 0 PCI-MSI-X eth-mlx4-11
84: 62 7482 71438 453805 67724 23715 109788 13430 636465 PCI-MSI-X eth1-0
90: 2 0 0 0 0 0 0 0 0 IO-APIC-level ehci_hcd:usb1, uhci_hcd:usb2
91: 0 0 0 0 0 0 0 0 0 PCI-MSI-X eth-mlx4-12
92: 24 143389 26631 554 21221 2572 15724 5514 260811 PCI-MSI-X eth1-1
98: 42 0 0 0 0 0 0 0 0 IO-APIC-level uhci_hcd:usb3, uhci_hcd:usb5
99: 0 0 0 0 0 0 0 0 0 PCI-MSI-X eth-mlx4-13
100: 21 2873 16742 199257 19787 1764 16691 18223 260311 PCI-MSI-X eth1-2
106: 39 0 0 2 0 0 0 0 0 IO-APIC-level uhci_hcd:usb4, uhci_hcd:usb6
107: 0 0 0 0 0 0 0 0 0 PCI-MSI-X eth-mlx4-14
108: 21 6813 171122 11072 5053 666 9692 8617 574229 PCI-MSI-X eth1-3
115: 0 0 0 0 0 0 0 0 0 PCI-MSI-X eth-mlx4-15
116: 24 6629 7046 3427 16290 260 7102 24153 1289959 PCI-MSI-X eth1-4
123: 0 0 0 0 0 0 0 0 0 PCI-MSI-X eth-mlx4-16
124: 36 4323 6746 831 8212 8321 27072 1072 119749 PCI-MSI-X eth1-5
131: 0 0 0 0 0 0 0 0 0 PCI-MSI-X eth-mlx4-17
132: 33 852 16528 10581 24815 786 12471 3345 546 PCI-MSI-X eth1-6
138: 7530 48097 277143 20863 353226 8399 112676 2274 0 PCI-MSI-X cciss0
139: 0 0 0 0 0 0 0 0 0 PCI-MSI-X eth-mlx4-18
140: 21 5237 23011 241674 38028 904 53358 143952 1360898 PCI-MSI-X eth1-7
147: 0 0 0 0 0 0 0 0 0 PCI-MSI-X eth-mlx4-19
154: 223 679759 809041 259138 346892 67788 41272 88795 0 IO-APIC-level ata_piix
155: 0 0 0 0 0 0 0 0 0 PCI-MSI-X eth-mlx4-20
163: 0 0 0 0 0 0 0 0 0 PCI-MSI-X eth-mlx4-21
171: 0 0 0 0 0 0 0 0 0 PCI-MSI-X eth-mlx4-22
179: 0 0 0 0 0 0 0 0 0 PCI-MSI-X eth-mlx4-23
186: 1233 72090 145452 398645 300622 15190867 50411 149256 16761877 PCI-MSI-X eth-mlx4-0
187: 0 0 0 0 0 0 0 0 0 PCI-MSI-X eth-mlx4-24
194: 0 0 0 0 0 0 0 0 0 PCI-MSI-X eth-mlx4-1
195: 2793 860 0 0 0 0 0 0 0 PCI-MSI-X mlx4_core(async)
202: 0 0 0 0 0 0 0 0 0 PCI-MSI-X eth-mlx4-2
203: 1085 552266 1638 8305 0 91290 0 10453 2445009 PCI-MSI-X eth0-0
210: 0 0 0 0 0 0 0 0 0 PCI-MSI-X eth-mlx4-3
211: 21 75686 221941 153330 139415 64270 43745 12065 23967 PCI-MSI-X eth0-1
218: 0 0 0 0 0 0 0 0 0 PCI-MSI-X eth-mlx4-4
219: 51 53715 155960 49802 160023 72700 423114 89175 14579 PCI-MSI-X eth0-2
226: 0 0 0 0 0 0 0 0 0 PCI-MSI-X eth-mlx4-5
227: 81 105457 349192 82962 312744 65128 174255 61067 23275 PCI-MSI-X eth0-3
234: 0 0 0 0 0 0 0 0 0 PCI-MSI-X eth-mlx4-6
235: 52 79276 223699 148956 373883 69912 114500 55452 22314 PCI-MSI-X eth0-4
NMI: 25485 25951 25333 23303 15267 23768 24350 28033
LOC: 209451763 209451686 209451617 209451543 209451472 209451400 209451328 209451256 209450670
Linux 2.6.18

edit out-matrosov-interupts - Far 3.0.2927 x86

C:\Users\KONSTA~1\AppData\Local\Temp\F3T2545.tmp\out-matrosov-interupts

Thu Sep 12 08:51:45 NOVT 2013

	CPU0	CPU1	CPU2	CPU3	CPU4	CPU5	CPU6	CPU7		
0:	45	0	0	0	0	0	0	0	IO-APIC-edge	timer
1:	1	422229	0	0	0	0	0	0	IO-APIC-edge	i8042
8:	0	1	0	0	0	0	0	0	IO-APIC-edge	rtc0
9:	0	0	0	0	0	0	0	0	IO-APIC-fasteoi	acpi
16:	0	0	641	1	0	0	0	0	IO-APIC-fasteoi	ehci_hcd:usb1
17:	0	0	37395	34485	1251	14352	0	0	IO-APIC-fasteoi	firewire_ohci
23:	0	0	0	1199876	262551	1180414	829635	0	IO-APIC-fasteoi	ehci_hcd:usb2
40:	0	0	16717752	21218321	10582386	49404082	0	0	IO-APIC-fasteoi	nvidia
44:	0	0	0	0	982	0	0	0	IO-APIC-fasteoi	snd_hda_intel
64:	0	0	0	7232218	2397003	5830875	3670146	0	PCI-MSI-edge	ahci
65:	0	0	0	0	0	0	0	0	PCI-MSI-edge	ahci
66:	0	0	0	0	0	0	0	0	PCI-MSI-edge	ahci
67:	0	0	0	0	0	0	0	0	PCI-MSI-edge	xhci_hcd
68:	0	0	0	0	0	0	0	0	PCI-MSI-edge	xhci_hcd
69:	0	0	0	0	0	0	0	0	PCI-MSI-edge	xhci_hcd
70:	0	0	0	0	0	0	0	0	PCI-MSI-edge	xhci_hcd
71:	0	0	0	0	0	0	0	0	PCI-MSI-edge	xhci_hcd
72:	0	0	0	0	0	0	0	0	PCI-MSI-edge	xhci_hcd
73:	0	0	0	0	0	0	0	0	PCI-MSI-edge	xhci_hcd
74:	0	0	0	0	0	0	0	0	PCI-MSI-edge	xhci_hcd
75:	0	0	0	0	0	0	0	0	PCI-MSI-edge	xhci_hcd
76:	0	0	0	0	0	0	0	0	PCI-MSI-edge	xhci_hcd
77:	0	0	0	0	0	0	0	0	PCI-MSI-edge	xhci_hcd
78:	0	0	0	0	0	0	0	0	PCI-MSI-edge	xhci_hcd
79:	0	0	0	0	0	0	0	0	PCI-MSI-edge	xhci_hcd
80:	0	0	0	0	0	0	0	0	PCI-MSI-edge	xhci_hcd
81:	0	0	0	0	0	0	0	0	PCI-MSI-edge	xhci_hcd
82:	0	0	0	0	0	0	0	0	PCI-MSI-edge	xhci_hcd
83:	0	0	0	9	0	0	0	0	PCI-MSI-edge	mei
84:	0	0	0	0	957	0	0	0	PCI-MSI-edge	snd_hda_intel
85:	0	0	0	0	13824584	0	0	0	PCI-MSI-edge	eth0-0
86:	0	6128648	0	0	7	0	0	0	PCI-MSI-edge	eth0-1
87:	0	0	13591556	0	0	24	0	0	PCI-MSI-edge	eth0-2
88:	0	0	0	5243378	0	3	0	0	PCI-MSI-edge	eth0-3
89:	6161120	0	0	0	0	7	0	0	PCI-MSI-edge	eth0-4
NMI:	1	1	1	3	0	1	1	0	Non-maskable interrupts	
LOC:	535381849	474755235	500266121	464582437	51359907	35686540	86207209	60270326	Local timer interrupts	
SPU:	0	0	0	0	0	0	0	0	Spurious interrupts	
PMI:	1	1	1	3	0	1	1	0	Performance monitoring interrupts	
IWI:	0	0	0	0	0	0	0	0	IRQ work interrupts	
RTR:	7	0	0	0	0	0	0	0	APIC ICR read retries	
RES:	205124573	61884458	1603405	174437	115150	79936	148497	78809	Rescheduling interrupts	
CAL:	1988733	2034833	1712724	1776644	2126671	1880745	2165410	2171341	Function call interrupts	
TLB:	3353025	3022253	3179024	2965854	236014	249937	274848	272985	TLB shootdowns	
TRM:	0	0	0	0	0	0	0	0	Thermal event interrupts	
THR:	1	1	1	1	1	1	1	1	Threshold APIC interrupts	
MCE:	0	0	0	0	0	0	0	0	Machine ch	
MCP:	18665	18665	18665	18665	18665	18665	18665	18665	Machine ch	
ERR:	0									

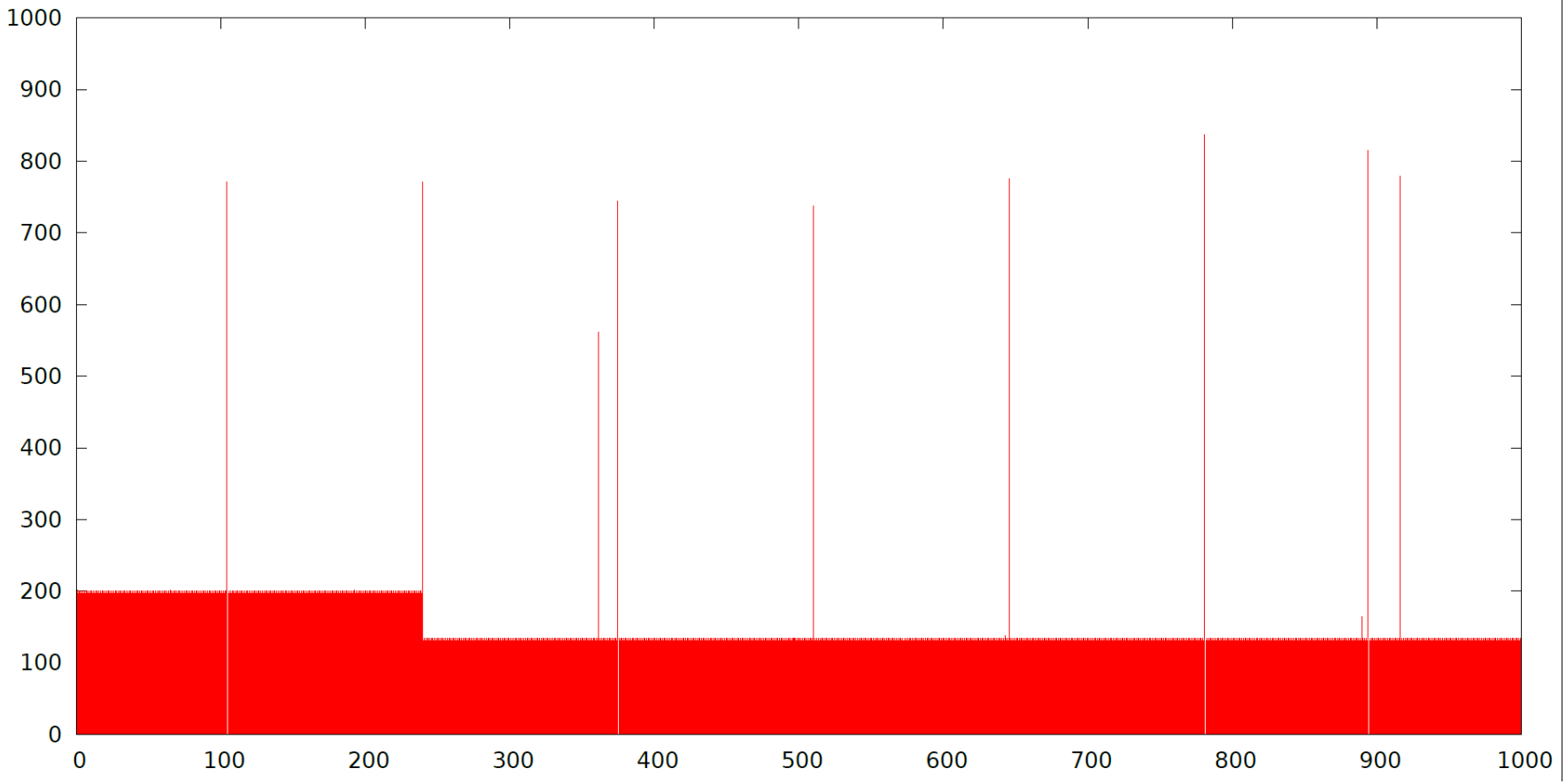
Linux 2.6.32

Измерение прерываний

Длительность и частота системных прерываний определялись следующим образом:

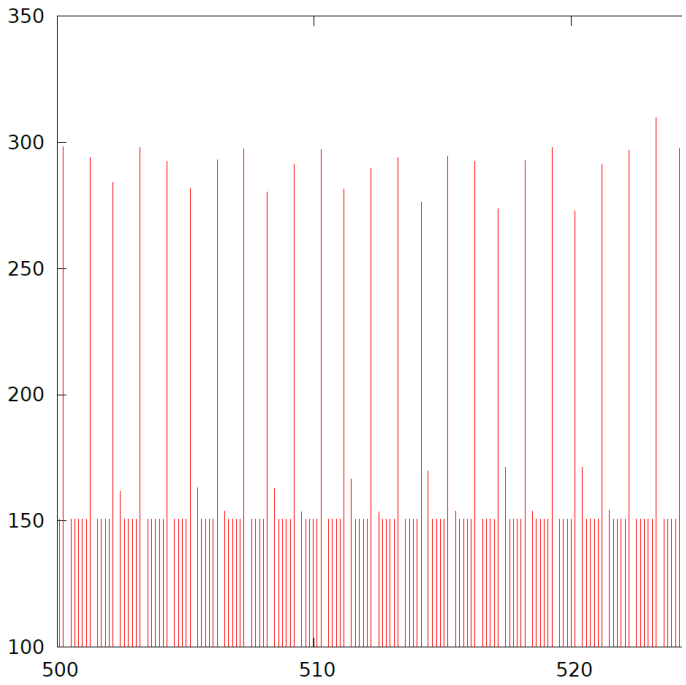
- В цикле 10^5 раз запускалась счётная функция, вычисляющая числа Фибоначчи, работающая в среднем T_C мкс
- Длительность исполнения каждого такого запуска сохранялась в памяти, и по завершении цикла записывалась в файл.

Прерывания

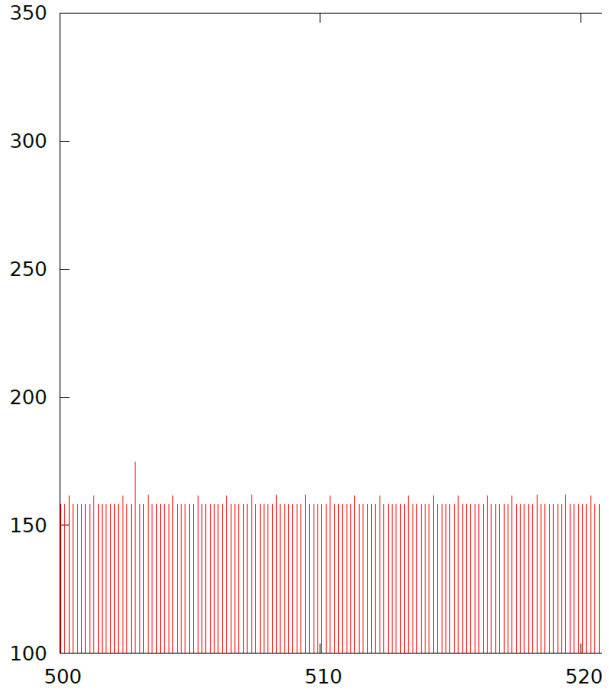


Не более первой четверти секунды работы программы время работы счётной функции в 1.5-2 раза больше среднего

Прерывания



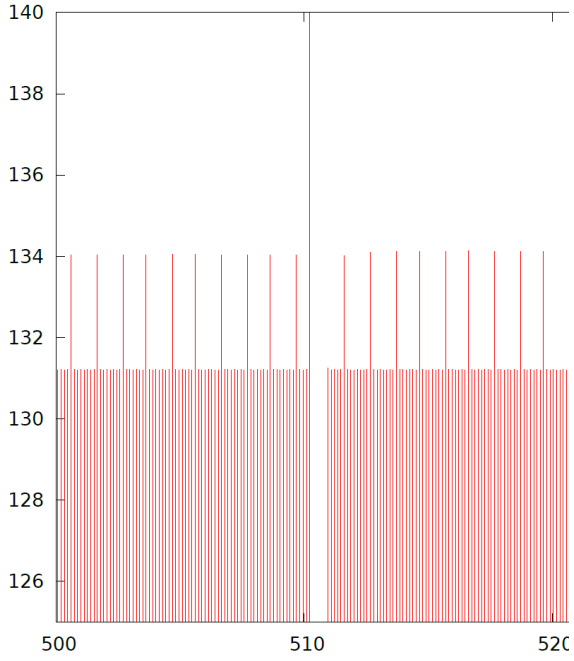
psize = 1
core = 0



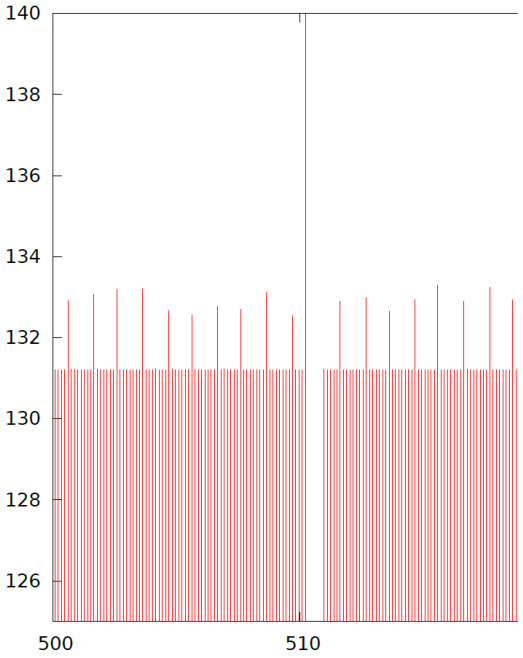
psize = 8
core = 0

Длительность обработки прерываний зависит от загрузки узла

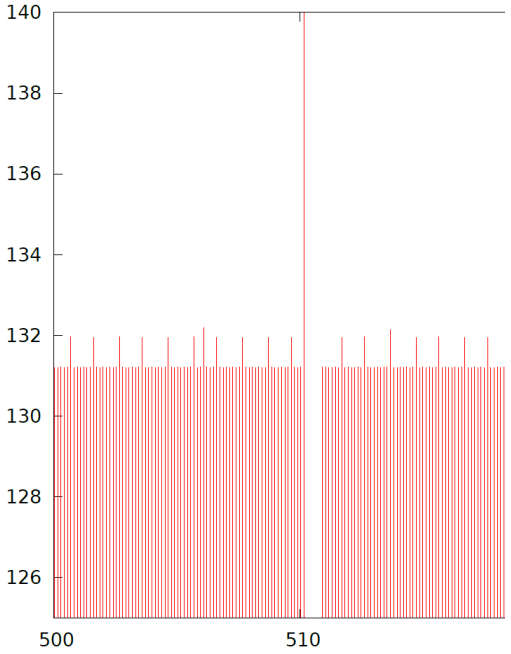
Прерывания



psize = 8
core = 0



psize = 8
core = 1



psize = 8
core = 4

Время обработки прерываний на 0-м ядре больше, чем на остальных ядрах

Длительность прерываний

Частота

Длительность мкс

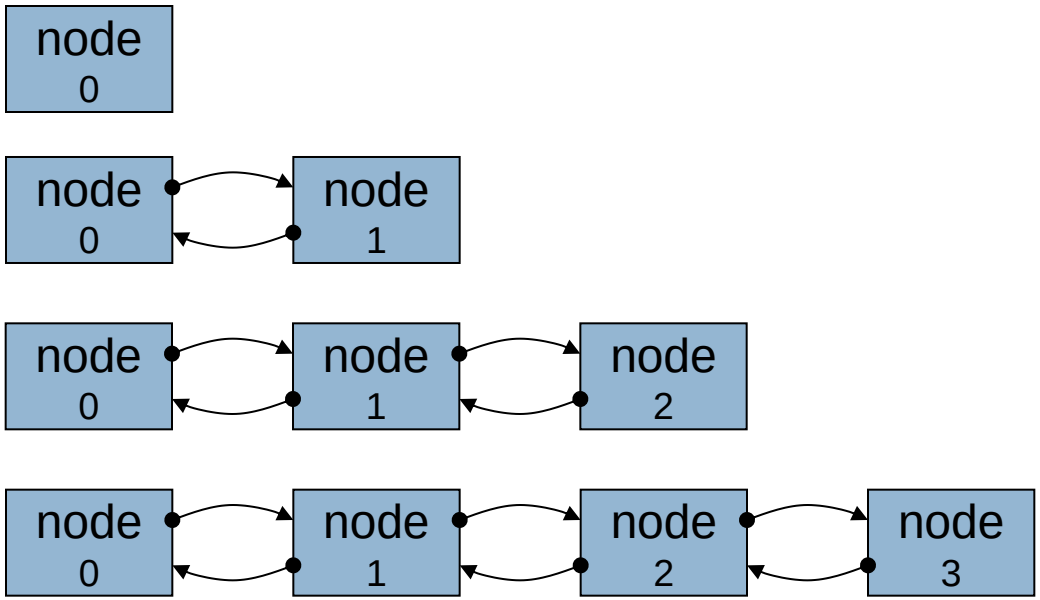
	Hz	одно ядро ядро 0	одно ядро ядро 1	Hz	весь узел ядро 0	весь узел ядро 1
НКС-30Т g6/g7	1000	90	13		3	1
	8	160	160		90	90
МВС 100К	1000	3			1	
	8	580			580	
МВС 10П	1000	2		100	2	
	40	7		10	7	
	25	13		2	13	
	12	25		0.5	50	

Много или мало?

Параллельная программа

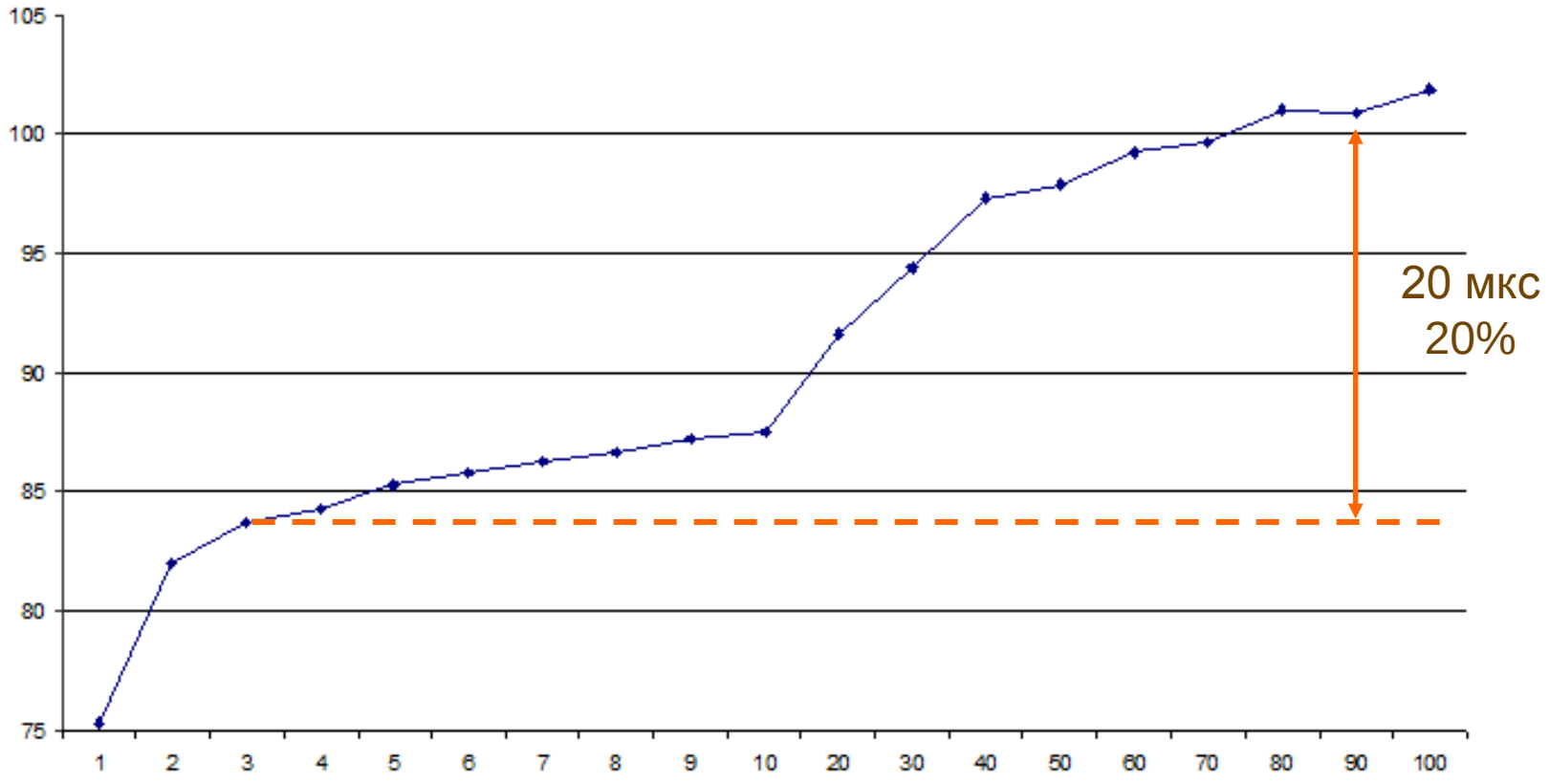
На каждой итерации предыдущей программы происходит обмен значениями между «соседними» процессами (1024 байт)

Время работы без прерываний



1. compute
2. MPI_Isend + MPI_Irecv + MPI_waitall

Время работы (МСЦ 100К)

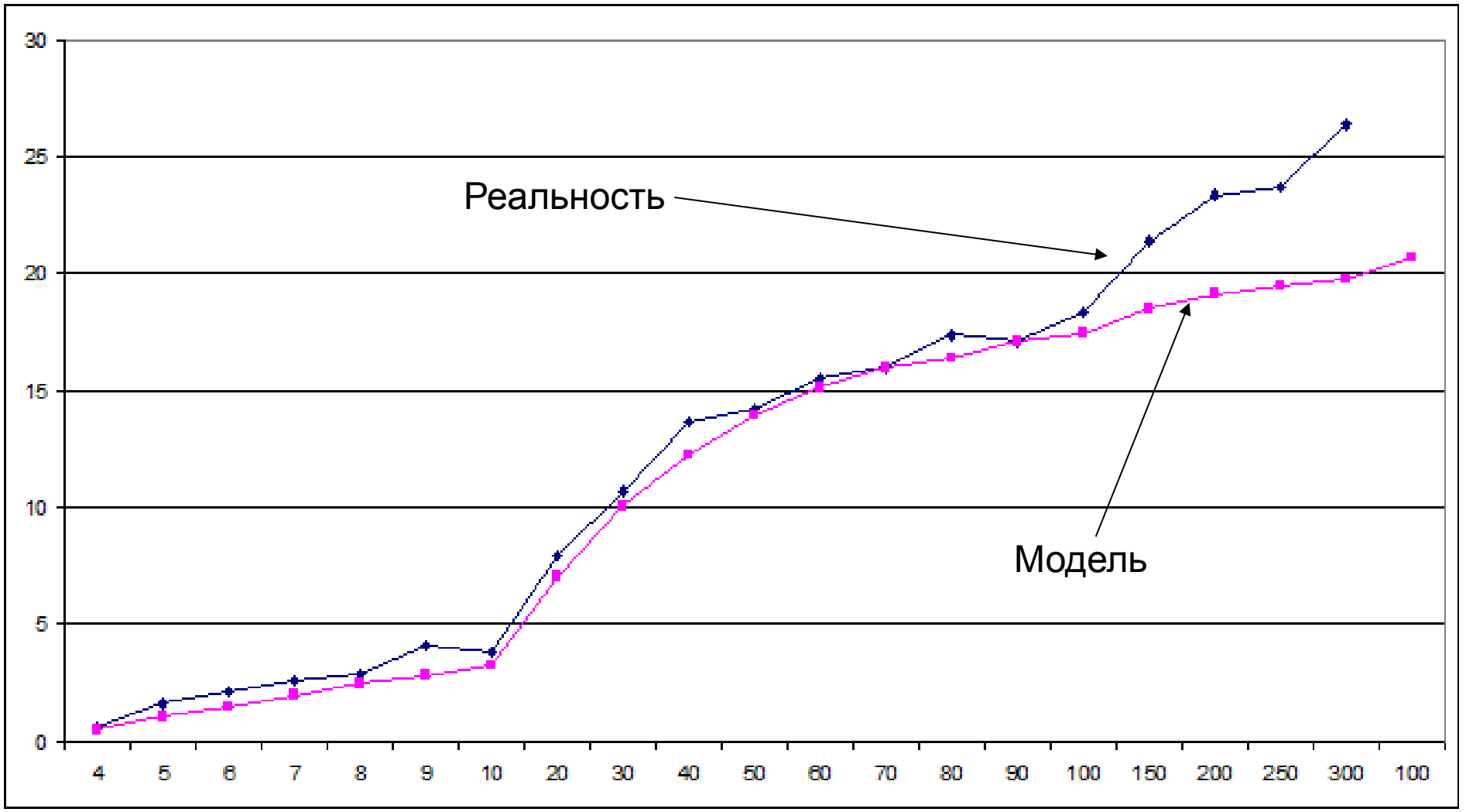


Модель распространения задержек

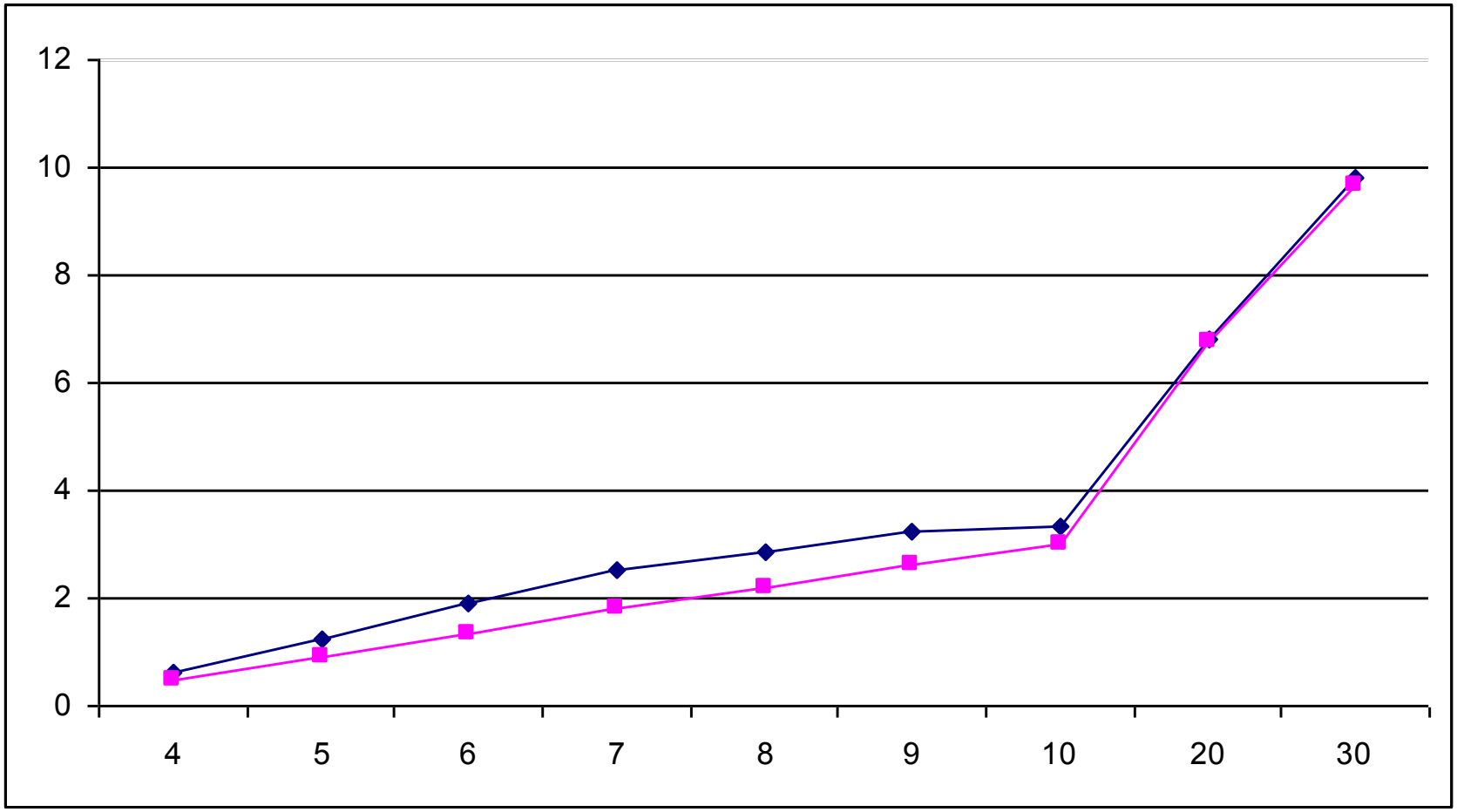
Суммарное время вычисления $i+1$ итерации процессом k, t_k^i, определяется следующим образом:

$$
\begin{aligned}
t_k^0 &= 0, \quad \forall 1 \le k \le P \\
t_1^{i+1} &= \max\left(t_1^i, t_2^i\right) + T_C + r_1^{i+1} \\
t_k^{i+1} &= \max\left(t_{k-1}^i, t_k^i, t_{k+1}^i\right) + T_C + r_k^{i+1}, \quad 1 < k < p \\
t_p^{i+1} &= \max\left(t_{p-1}^i, t_p^i\right) + T_C + r_p^{i+1}
\end{aligned}
$$

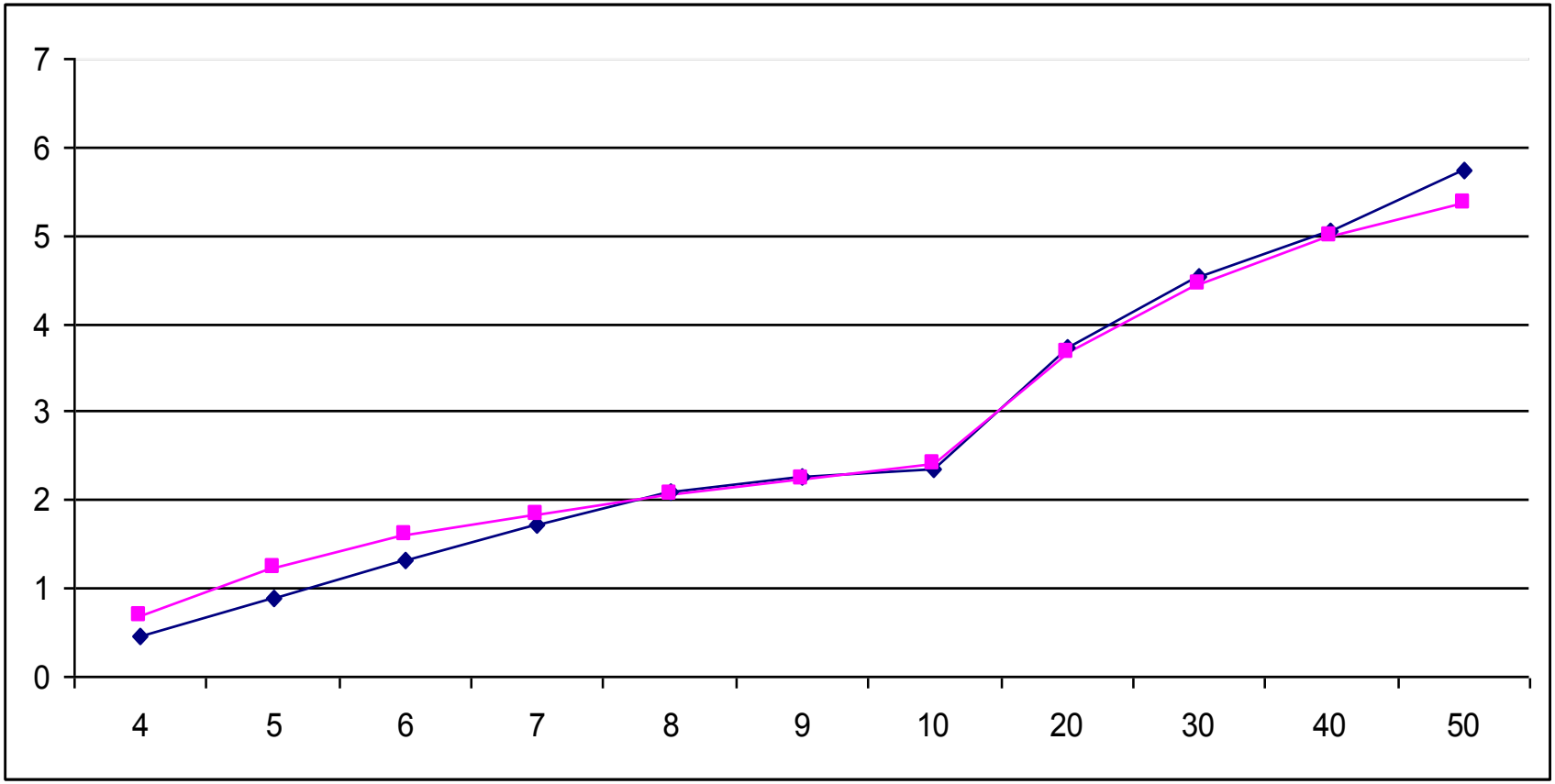
МСЦ 100К : T_P-T_3



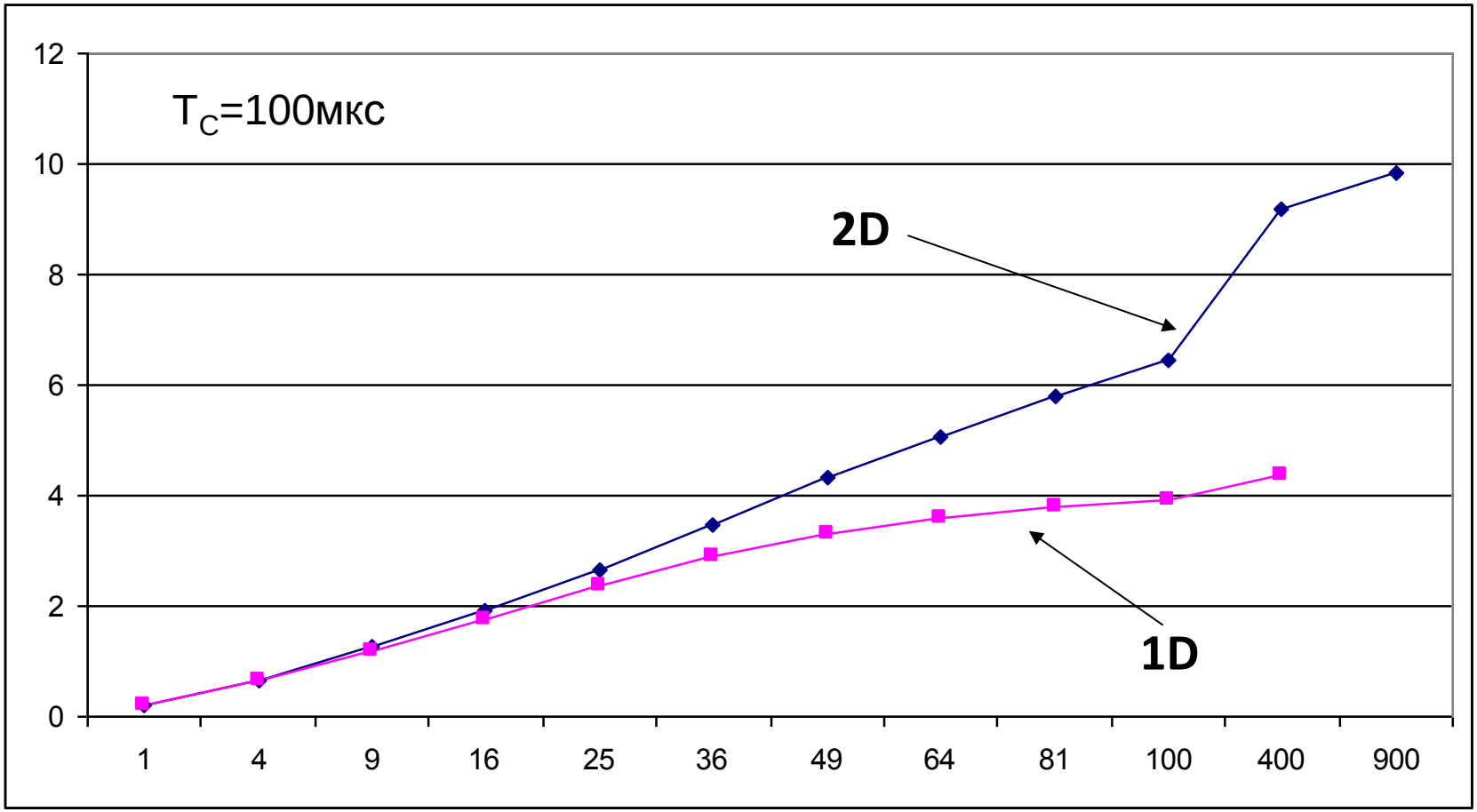
НКС-30 T, workq, T_C=77мкс



НКС-30 Т, g6, T_C=82мкс



Моделирование 2D



Заключение

- Построена модель распространения задержек, связанных с системными прерываниями, в системе параллельно работающих процессов. Показана её адекватность.
- Показано, что прерывания являются одной из причин деградации производительности параллельного исполнения с ростом числа используемых процессоров.
- Влияние системных прерываний может быть больше времени коммуникации процессов в разы.
- В 2D и в 3D декомпозициях области влияние системных прерываний увеличивается (по сравнению с 1D).
- Необходимо:
  - (a) использовать **неблокируемые** способы коммуникации, которые будут препятствовать излишней синхронизации процессов
  - (b) внедрять **ОС с минимальным «системным шумом»**
  - (c) тестирование на определение масштабируемости, производительности, ускорения или эффективности параллельной реализации проводить лишь **после четверти секунды** работы этих ядер под нагрузкой
  - Также необходимо учитывать изменения частоты ядер от загруженности узла в целом