

Методы параллельной обработки сверхбольших баз данных с использованием распределенных колоночных индексов

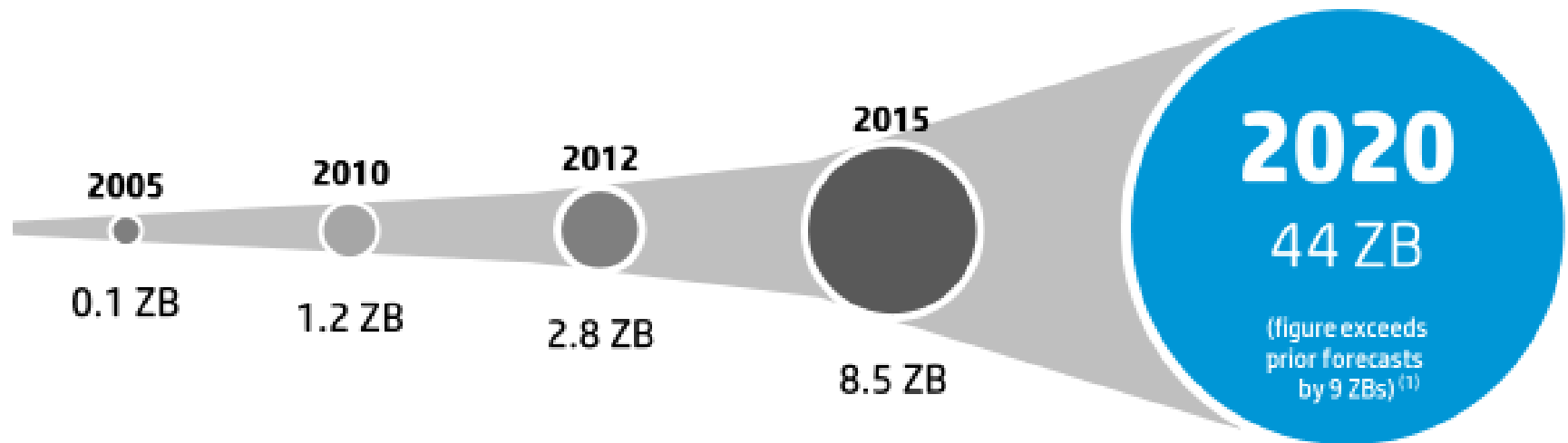
05.13.11 - математическое и программное обеспечение
вычислительных машин, комплексов и компьютерных сетей

Диссертация на соискание ученой степени кандидата физико-математических наук

Елена Владимировна Иванова

Научный руководитель:
СОКОЛИНСКИЙ Леонид Борисович,
доктор физ.-мат. наук, профессор

Цифровая вселенная



- Объем хранимой информации удваивается каждые два года
- из всего объема существующих данных потенциально полезны 22%, из которых менее 5% были подвергнуты анализу

Тенденции в развитии технологий обработки больших данных

- Для обработки больших данных необходимо использовать СУБД
(Майкл Стоунбрейкер)
- Кластерные вычислительные системы с большой суммарной оперативной памятью
- Базы данных в оперативной памяти
- Многоядерные ускорители
- Колоночное представление данных со сжатием

Цель диссертационной работы

Разработка и исследование эффективных методов параллельной обработки сверхбольших баз данных с использованием колоночного представления информации, ориентированных на кластерные вычислительные системы, оснащенные многоядерными ускорителями, и допускающих интеграцию с реляционными СУБД

Основные задачи

1. Разработать колоночные индексы и методы их фрагментации
2. Разработать методы декомпозиции реляционных операций для распределенных колоночных индексов
3. Реализовать предложенные подходы и методы в виде колоночного сопроцессора КСОП для кластерных вычислительных систем
4. Провести вычислительные эксперименты с использованием КСОП

Колоночное и строчное представление данных

Sales				
	saleid	prodid	date	region
1				
2				
3				
4				
5				
6				
7				
8				
9				
10				

Колоночная модель
с виртуальными ключами

Sales				
	saleid	prodid	date	region
1		1		1
2		2		2
3		3		3
4		4		4
5		5		5
6		6		6
7		7		7
8		8		8
9		9		9
10		10		10

Колоночная модель
с суррогатными ключами

Sales				
	saleid	prodid	date	region
1				
2				
3				
4				
5				
6				
7				
8				
9				
10				

Строчная модель

Преимущества и недостатки колоночного представления данных

Преимущества

Колоночные СУБД в 10 раз эффективнее строчных на рабочих нагрузках, характерных для хранилищ данных

- Колоночное представление позволяет загружать только те столбцы, которые используются в запросах
- Колоночное представление позволяет более эффективно использовать:
 - кэш-память процессора
 - команды векторной обработки
 - предварительную выборку данных
 - сжатие данных

Недостатки

- Колоночные СУБД проигрывают строчным на OLTP-запросах
- Для колоночных СУБД не применима техника оптимизации, основанная на реляционной алгебре

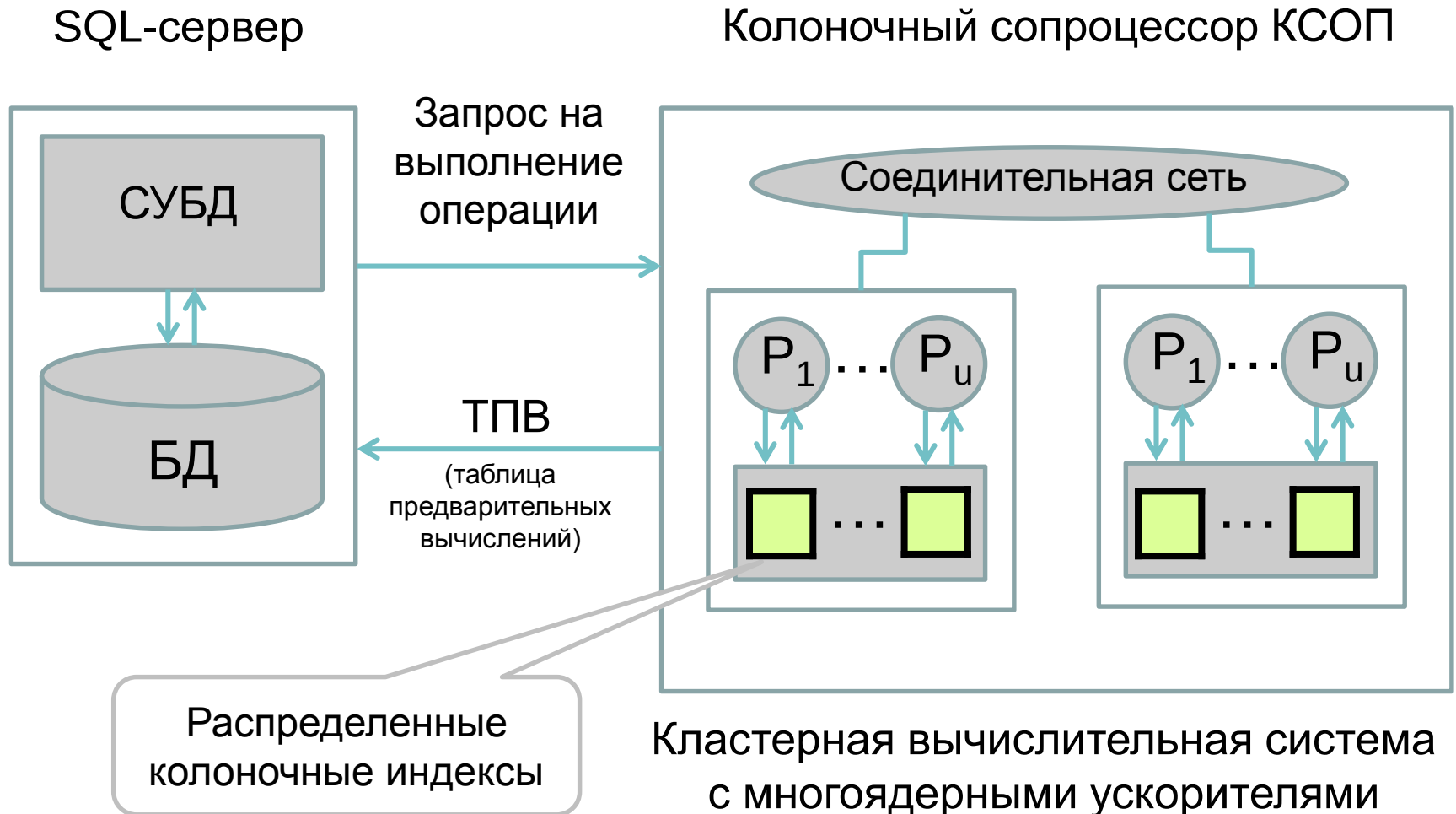
Исследовательская задача

Соединить в одной системе
преимущества колоночного и строчного
хранения данных

Работы по теме диссертации

Ramamurthy R., Dewitt D., Su Q. A case for fractured mirrors // Proceedings of the VLDB Endowment. 2002. Vol. 12, No. 2. P. 89-101.	«Разбитое зеркало»
Abadi D.J., Madden S.R., Hachem N. Column-Stores vs. Row-Stores: How Different Are They Really? // Proceedings of the 2008 ACM SIGMOD international conference on Management of data, June 9-12, 2008, Vancouver, BC, Canada. ACM, 2008. P. 967-980.	Эмуляция колоночного представления в строчной СУБД
Bruno N. Teaching an Old Elephant New Tricks // Online Proceedings of Fourth Biennial Conference on Innovative Data Systems Research (CIDR 2009), Asilomar, CA, USA, January 4-7, 2009.	С-таблицы
El-Helw A., Ross K.A., Bhattacharjee B., Lang C.A., Mihaila G.A. Column-oriented query processing for row stores // Proceedings of the ACM 14th international workshop on Data Warehousing and OLAP (DOLAP '11), October 28, 2011, Glasgow, United Kingdom. ACM, 2011. P. 67-74.	Только индексные планы
Larson P.-A., Clinciu C., Hanson E. N., Oks A., Price S. L., Rangarajan S., Surna A., Zhou Q. SQL server column store indexes // Proceedings of the 2011 ACM SIGMOD International Conference on Management of data (SIGMOD '11), June 12-16, 2011, Athens, Greece. ACM, 2011. P. 1177-1184.	Индексы колоночной памяти (column store indexes)
Jha S., He B., Lu M., Cheng X., Huynh H. P. Improving main memory hash joins on Intel Xeon Phi processors: an experimental approach // Proceedings of the VLDB Endowment. 2015. Vol. 8, No. 6. P. 642-653.	Использование Xeon Phi

Предлагаемое решение



Колоночный индекс

$I_{R.B}$
(колоночный индекс
для B)

A	B
3	110
1	114
5	127
0	136
2	136
7	158
4	174
6	174
8	187

Таблица R

A	B	C
0	136	17
1	114	10
2	136	25
3	110	10
4	174	15
5	127	99
6	174	97
7	158	63
8	187	55

$I_{R.C}$
(колоночный индекс
для C)

A	C
1	10
3	10
4	15
0	17
2	25
8	55
7	63
6	97
5	99

Формальное определение колоночного индекса

Пусть $R(A, B, \dots)$ – отношение R с суррогатным ключом A и атрибутом B . A состоит из целочисленных неотрицательных элементов.

$\mathfrak{D}_B$ – домен атрибута B . На множестве $\mathfrak{D}_B$ задано отношение линейного порядка. $T(R) = n$ – количество элементов в R .

Колоночным индексом $I_{R.B}$ атрибута B отношения R называется упорядоченное отношение, удовлетворяющее следующим требованиям:

$$T(I_{R.B}) = n \quad \text{и} \quad \pi_A(I_{R.B}) = \pi_A(R);$$

$$\forall x_1, x_2 \in I_{R.B} \left(x_1 \leq x_2 \Leftrightarrow x_1.B \leq x_2.B \right);$$

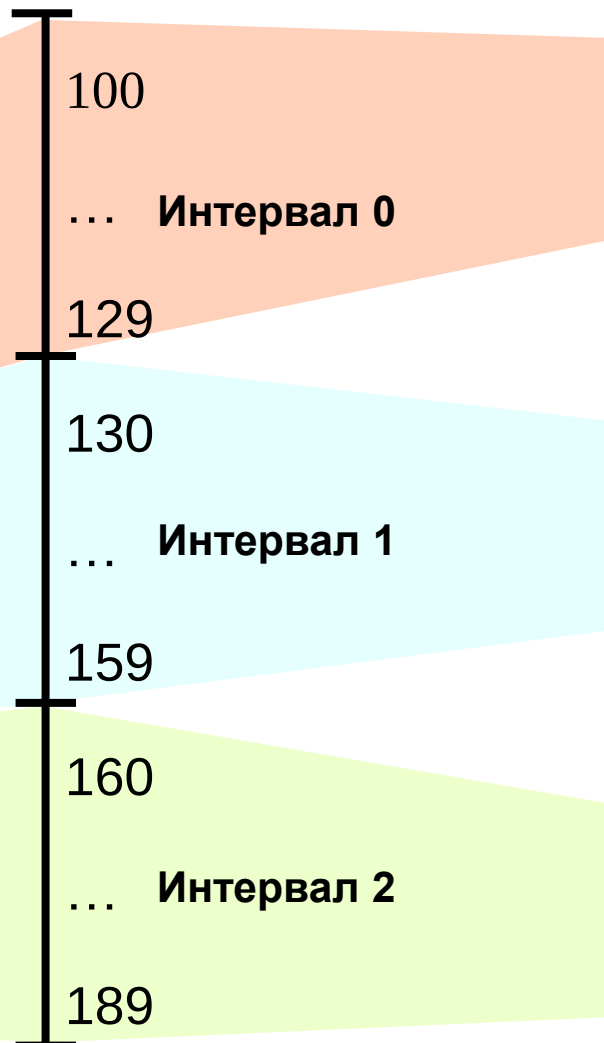
$$\forall r \in R \left(\forall x \in I_{R.B} \left(r.A = x.A \Rightarrow r.B = x.B \right) \right).$$

Доменно-интервальная фрагментация

Колоночный
индекс

A	B
3	110
1	114
5	127
0	136
2	136
7	158
4	174
6	174
8	187

Домен B



A	B
3	110
1	114
5	127

Фрагмент 0

A	B
3	136
1	136
5	158

Фрагмент 1

A	B
3	174
1	174
5	187

Фрагмент 2

Формальное определение доменно-интервальной фрагментации

Разбиение домена $\mathfrak{D}_B$ на k непересекающихся интервалов:

$$V_0 = [v_0; v_1); V_1 = [v_1; v_2); \dots; V_{k-1} = [v_{k-1}; v_k);$$

$$v_0 < v_1 < \dots < v_k; \quad \mathfrak{D}_B = \bigcup_{i=0}^{k-1} V_i$$

Доменная функция фрагментации

$$\varphi_{\mathfrak{D}_B}: \mathfrak{D}_B \rightarrow \{0, \dots, k-1\}$$

Доменно-интервальная функция фрагментации колоночного индекса: $\forall i \in \{0, \dots, k-1\} \left(\forall b \in \mathfrak{D}_B \left(\varphi_{\mathfrak{D}_B}(b) = i \Leftrightarrow b \in V_i \right) \right)$

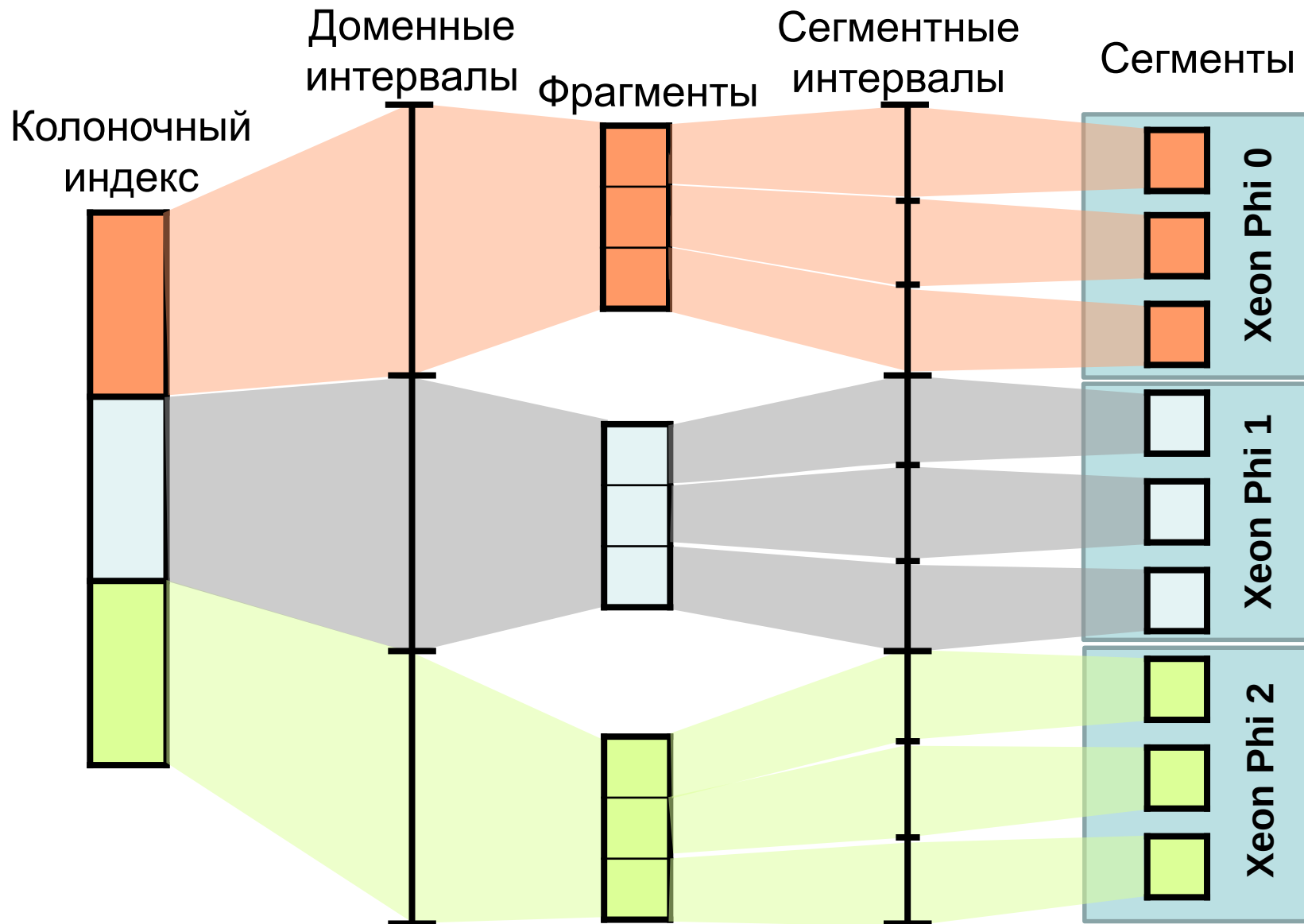
$$\varphi_{I_{R.B}}: I_{R.B} \rightarrow \{0, \dots, k-1\}$$

$$\forall x \in I_{R.B} \left(\varphi_{I_{R.B}}(x) = \varphi_{\mathfrak{D}_B}(x.B) \right)$$

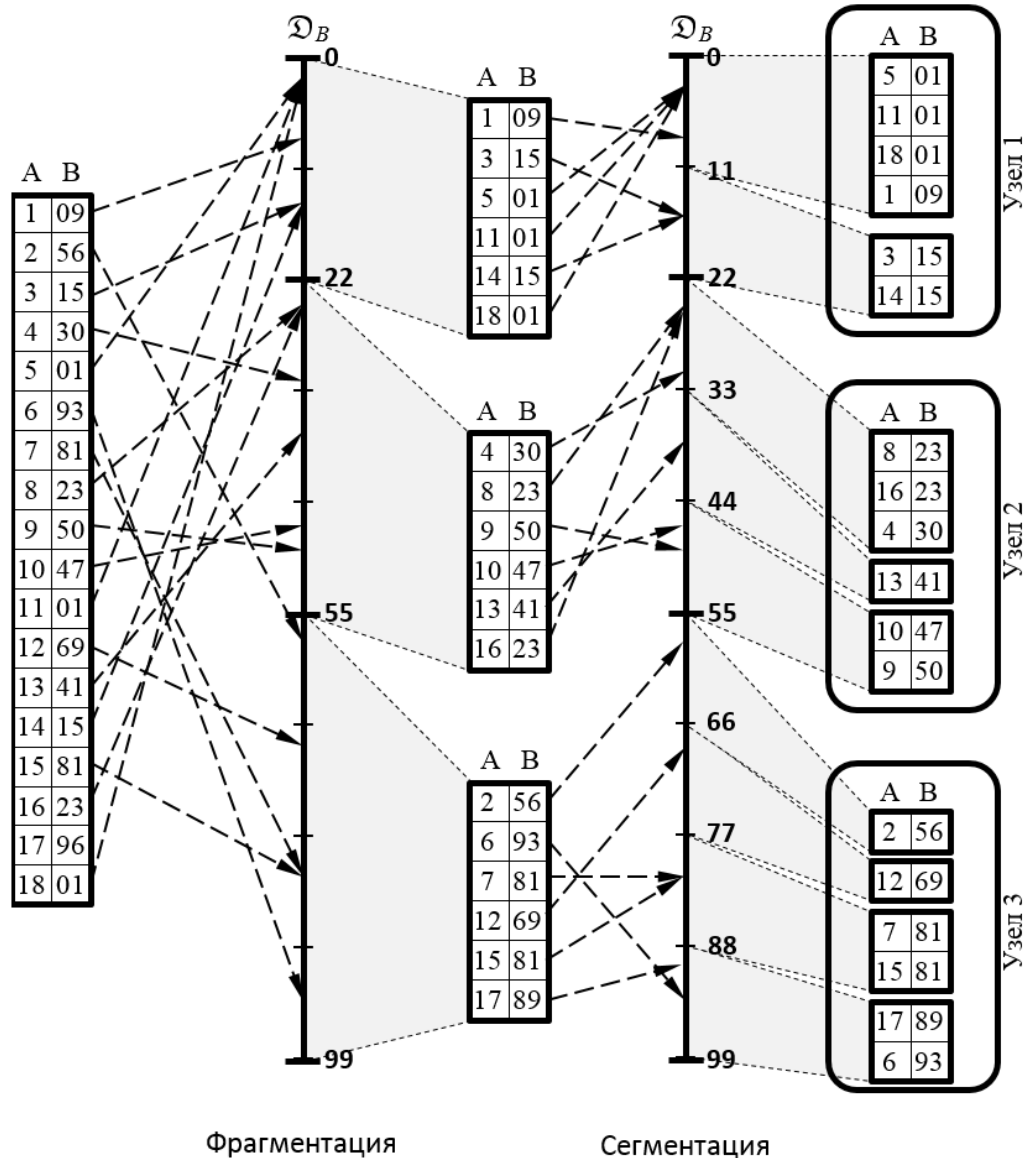
i -тый фрагмент колоночного индекса:

$$I_{R.B}^i = \left\{ x \mid x \in I_{R.B}; \varphi_{I_{R.B}}(x) = i \right\}$$

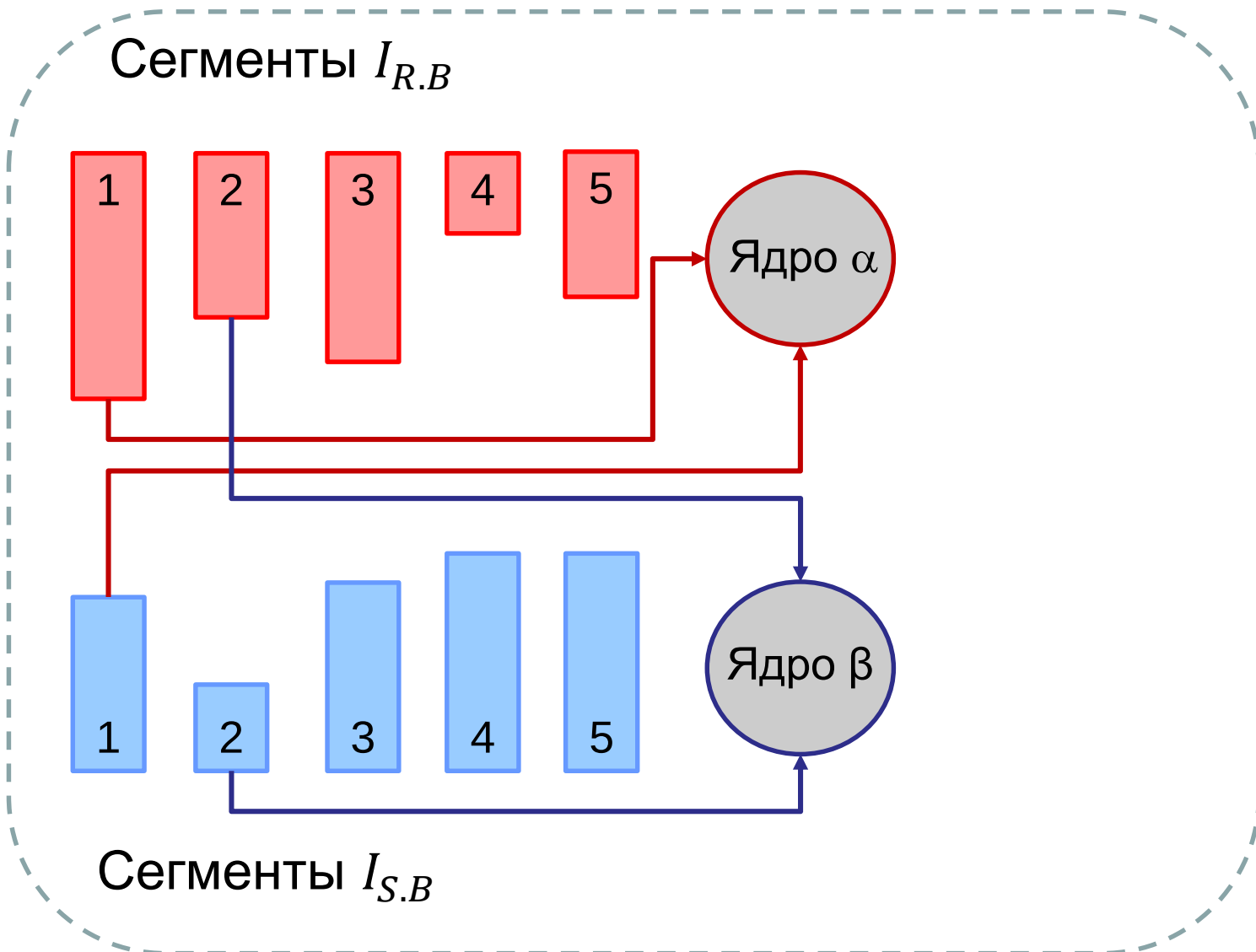
Использование многоядерных ускорителей Intel Xeon Phi



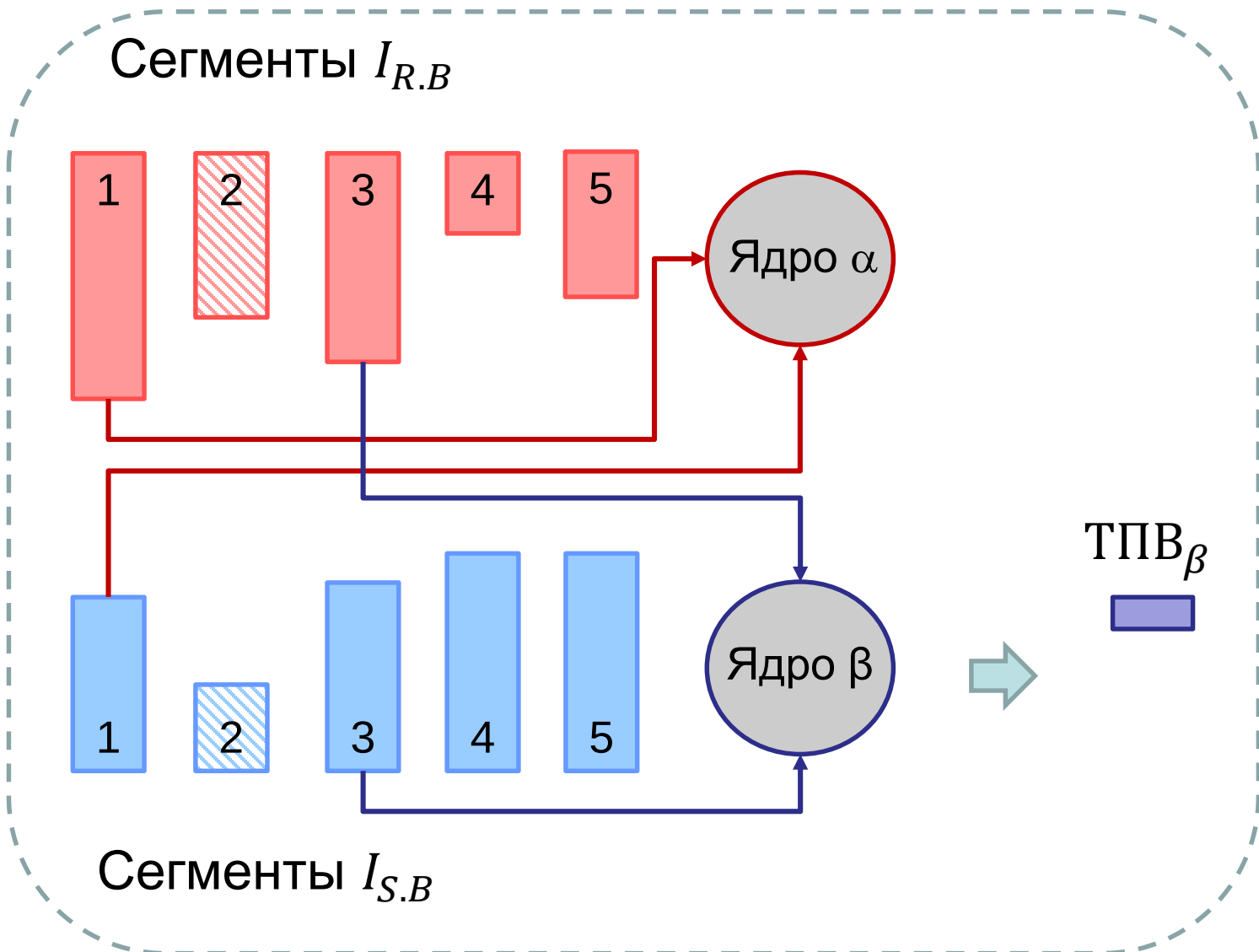
Пример двухуровневого разбиения колоночного индекса на фрагменты и сегменты



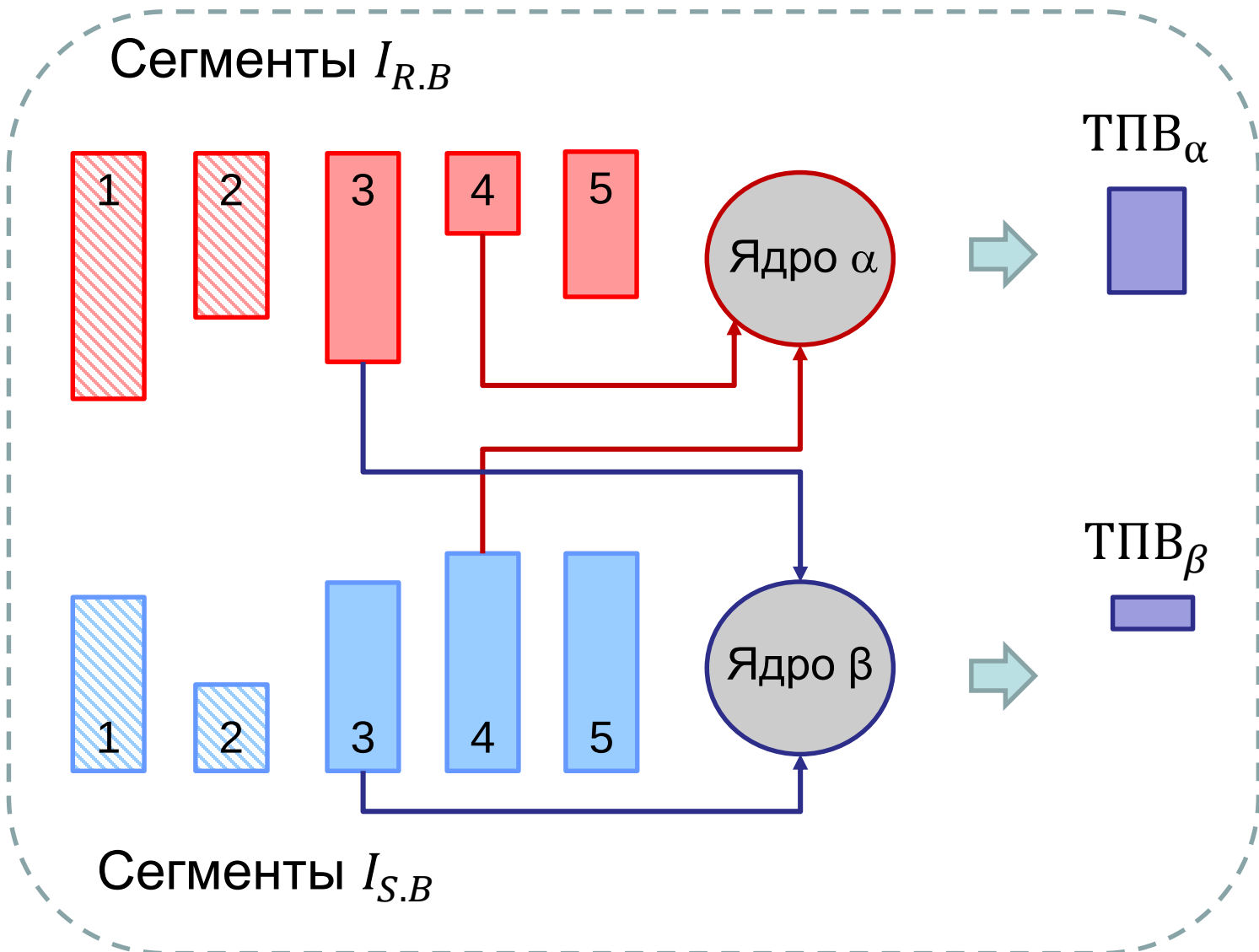
Балансировка загрузки ядер



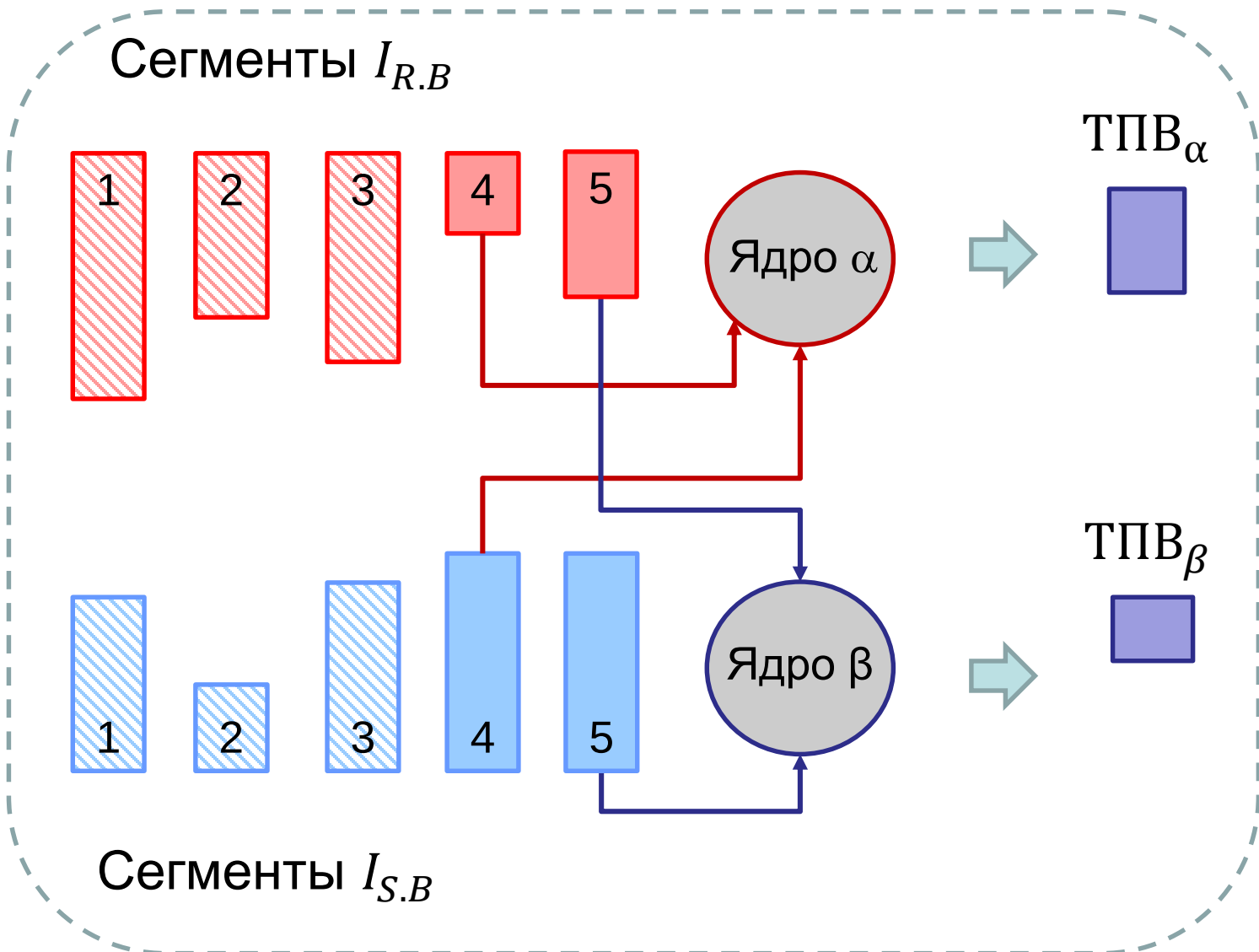
Балансировка загрузки ядер



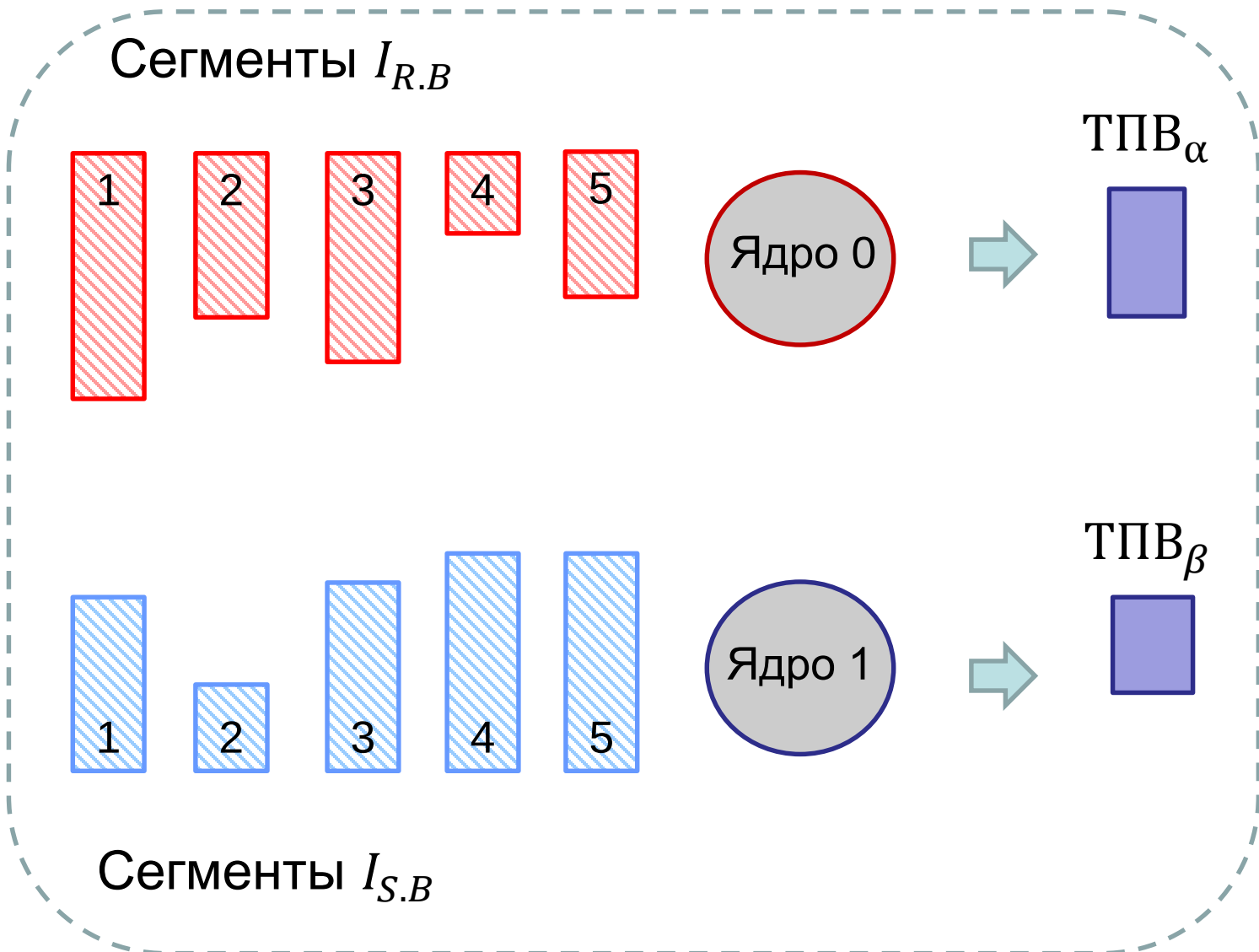
Балансировка загрузки ядер



Балансировка загрузки ядер



Балансировка загрузки ядер

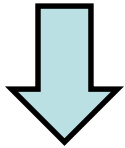


Параллельные алгоритмы выполнения реляционных операций

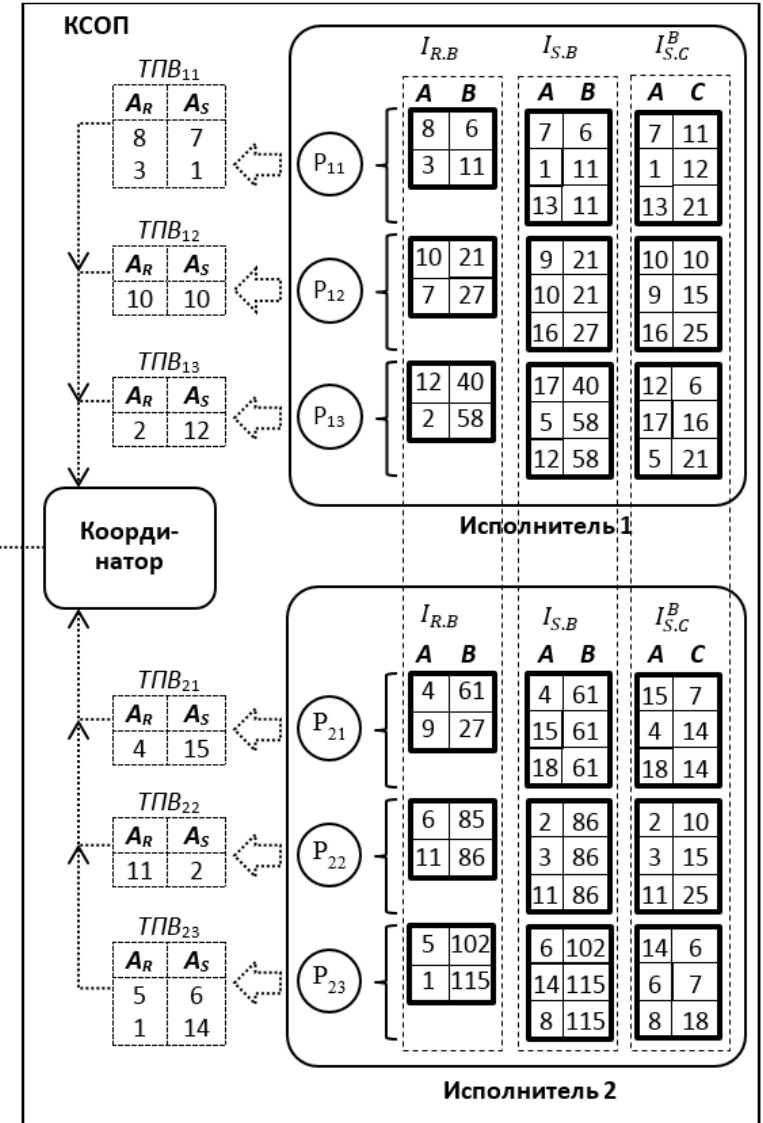
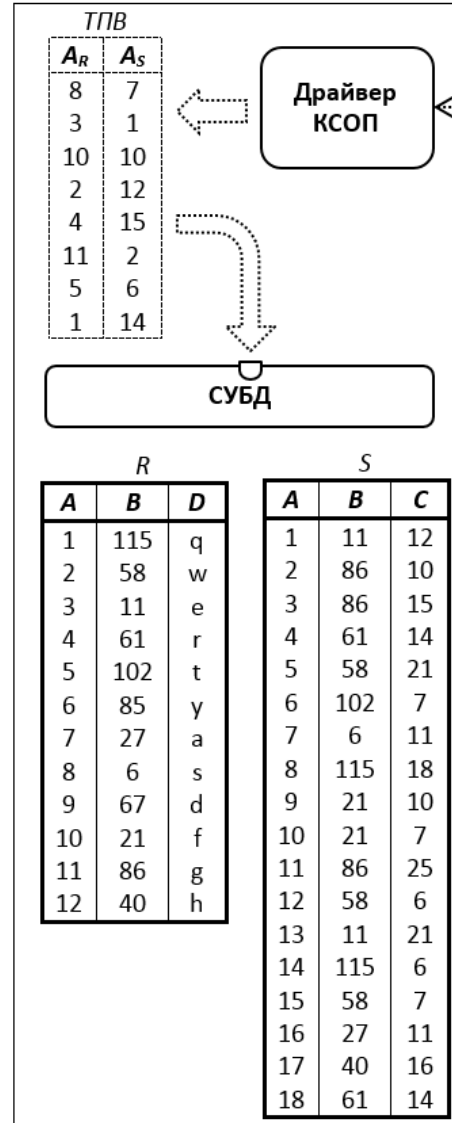
- Проекция
- Выбор
- Удаление дубликатов
- Пересечение
- Объединение
- Естественное соединение
- Группировка

Архитектуру системы баз данных с колоночным сопроцессором КСОП

```
SELECT D, C
FROM R, S
WHERE
  R.B = S.B
  AND C < 13;
```



$$\pi_{I_{R,B}.A \rightarrow A_R, I_{S,B}.A \rightarrow A_S} \left(I_{R,B} \bowtie_{\substack{I_{R,B}.B = \\ I_{S,B}.B}} \left(I_{S,B} \bowtie_{\sigma_{C < 13}} \left(I_{S,C}^B \right) \right) \right)$$



Выполнение запроса на SQL-сервере с помощью ТПВ

<i>R</i>		
<i>A</i>	<i>B</i>	<i>D</i>
1	115	q
2	58	w
3	11	e
4	61	r
5	102	t
6	85	y
7	27	a
8	6	s
9	67	d
10	21	f
11	86	g
12	40	h

<i>S</i>		
<i>A</i>	<i>B</i>	<i>C</i>
1	11	12
2	86	10
3	86	15
4	61	14
5	58	21
6	102	7
7	6	11
8	115	18
9	21	10
10	21	7
11	86	25
12	58	6
13	11	21
14	115	6
15	58	7
16	27	11
17	40	16
18	61	14

SELECT D, C
FROM

R INNER JOIN (
 ТПВ INNER JOIN S ON (S.A = ТПВ.A_S)
) ON (R.A = ТПВ.A_R)

<i>ТПВ</i>	
<i>A_R</i>	<i>A_S</i>
8	7
3	1
10	10
2	12
4	15
11	2
5	6
1	14



Результат	
<i>D</i>	<i>C</i>
s	11
e	12
f	7
w	6
r	7
g	10
t	7
q	6

Формальное определение декомпозиции естественного соединения $Q = \pi_{*\setminus A}(R) \bowtie \pi_{*\setminus A}(S)$

- Пусть $R(A, B_1, \dots, B_u, C_1, \dots, C_h)$ и $S(A, B_1, \dots, B_u, D_1, \dots, D_w)$.
- Пусть имеется два набора колоночных индексов по атрибутам $B_1, \dots, B_u$, для которых задана доменно-интервальная фрагментация степени k :

$$I_{R.B_j} = \bigcup_{i=0}^{k-1} I_{R.B_j}^i ; \quad I_{S.B_j} = \bigcup_{i=0}^{k-1} I_{S.B_j}^i .$$

- Положим для всех $i = 0, \dots, k-1$ и $j = 1, \dots, u$

$$P_j^i = \pi_{I_{R.B_j}^i.A \rightarrow A_R, I_{S.B_j}^i.A \rightarrow A_S} \left(I_{R.B_j}^i \bowtie_{I_{R.B_j}^i.B_j = I_{S.B_j}^i.B_j} I_{S.B_j}^i \right) \quad P_j = \bigcup_{i=0}^{k-1} P_j^i$$

$$P = \bigcap_{j=1}^u P_j \quad Q = \left\{ \left(\&_R(p.A_R).B_1, \dots, \&_R(p.A_R).B_u, \right. \right. \\ \&_B(p.A_B).C_1, \dots, \&_B(p.A_B).C_v, \\ \left. \&_S(p.A_S).D_1, \dots, \&_S(p.A_S).D_w \right) \mid p \in P \}.$$

Формальное определение транзитивной фрагментации

- Пусть для отношения $R(A, B, C, \dots)$ заданы колоночные индексы $I_{R.B}$ и $I_{R.C}$. *Транзитивной фрагментацией* индекса $I_{R.C}$ относительно индекса $I_{R.B}$ называется фрагментация, задаваемая функцией $\varphi_{I_{R.C}}^{I_{R.B}} : I_{R.C} \rightarrow \{0, \dots, k-1\}$, удовлетворяющей условию

$$\forall x \in I_{R.C} \left(\varphi_{I_{R.C}}^{I_{R.B}}(x) = \varphi_{I_{R.B}} \left(\sigma_{A=x.A} (I_{R.B}) \right) \right)$$

Реализация КСОП

- Язык программирования Си
- Технологии параллельного программирования MPI и OpenMP
- Исходные тексты приложения свободно доступны в сети Интернет:
<https://github.com/elena-ivanova/columnindices/>

Язык SSORQL

Код оператора (opcode)	Функция драйвера КСОП, порождающая оператор	Семантика оператора
1	CreateColumnIndex	Создание распределенного колоночного индекса
2	CreateTransitiveIndex	Создание транзитивно колоночного индекса
3	Execute	Выполнение запроса на вычисление ТПВ
4	Insert	Добавление кортежа в колоночный индекс
5	InsertBlock	Добавление блока кортежей в колоночный индекс
6	Update	Обновление значений кортежа в колоночном индексе
7	Delete	Удаление кортежа из колоночного индекса

Создание распределенного колоночного индекса

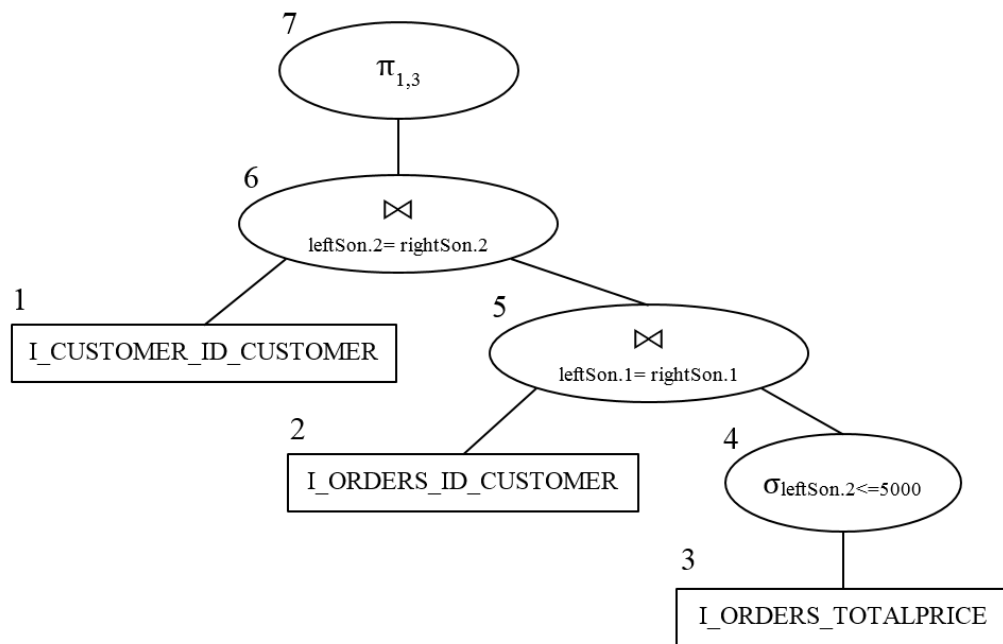
```
{  
  "opcode": 1,  
  "params": {  
    "CIndexID": 1,  
    "Width": 32,  
    "Bottom": 1,  
    "Top": 630000,  
    "Dimension": 1  
  }  
}
```

Вставка нового элемента в колоночный индекс

```
{  
  "opcode": 4,  
  "params": {  
    "CIndexID": 1,  
    "SurrogateKey": 0,  
    "Value": [38335];  
  }  
}
```

Оператор выполнения запроса на вычисление ТПВ

```
{  "opcode": 3,
  "queryPlan": [
    {  "nodeID": 1,
      "nodeType": "leaf",
      "indexID": I_CUSTOMER_ID_CUSTOMER    },
    {  "nodeID": 2,
      "nodeType": "leaf",
      "indexID": I_ORDERS_ID_CUSTOMER      },
    {  "nodeID": 3,
      "nodeType": "leaf",
      "indexID": I_ORDERS_TOTALPRICE      },
    {  "nodeID": 4,
      "nodeType": "inner",
      "leftSon": 3,
      "relOpCode": "selection",
      "parameters": "leftSon.2<=5000"    },
    {  "nodeID": 5,
      "nodeType": "inner",
      "leftSon": 2,
      "rightSon": 4,
      "relOpCode": "equijoin",
      "parameters": "leftSon.1= rightSon.1" },
    {  "nodeID": 6,
      "nodeType": "inner",
      "leftSon": 1,
      "rightSon": 5,
      "relOpCode": "equijoin",
      "parameters": "leftSon.2=rightSon.2" },
    {  "nodeID": 7,
      "nodeType": "root",
      "leftSon": 6,
      "relOpCode": "projection",
      "parameters": "1, 3"                }
  ]
}
```



Эксперименты

- Балансировка загрузки процессорных ядер Xeon Phi
- Влияние гиперпоточности
- Масштабируемость КСОП
- Использование КСОП при выполнении SQL-запросов

Характеристики вычислительных систем

	«Торнадо ЮУрГУ»	«RSC PetaStream»
Количество узлов:	384	64
Тип процессоров:	2 x Intel Xeon X5680 (12 ядер по 3.33 ГГц; 2 потока на ядро)	
Оперативная память узла:	24 Гб	
Тип сопроцессора:	Intel Xeon Phi SE10X: (61 ядро по 1.1 ГГц; 4 потока на ядро)	Intel Xeon Phi 7120 (61 ядро по 1.24 ГГц)
Память сопроцессора:	8 Гб	16 Гб
Тип системной сети:	InfiniBand QDR	InfiniBand FDR
Тип управляющей сети:	Gigabit Ethernet	Gigabit Ethernet
Операционная система:	Linux CentOS 6.2	Linux CentOS 7.0

Тестовая база данных

CUSTOMER
SF x 630 000

A
ID_CUSTOMER
NAME
ADDRESS
NATION
PHONE
ACCTBAL
MKTSEGMENT
COMMENT

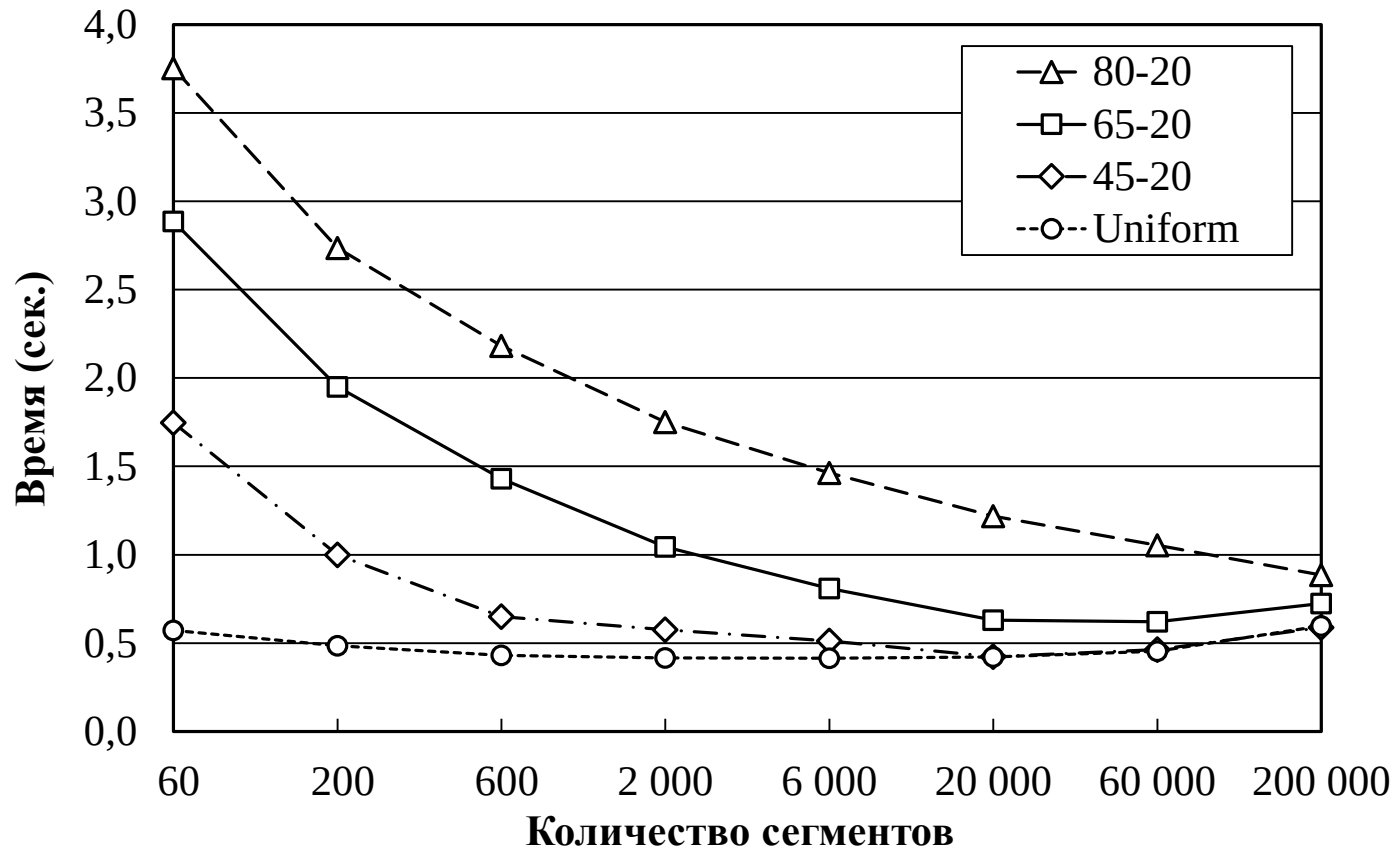
ORDERS
SF x 63 000 000

A
ID_ORDER
ID_CUSTOMER
LINENUMBER
ORDERSTATUS
TOTALPRICE
ORDERDATE
PRIORITY
CLERK
SHIPPRIORITY
...
COMMENT



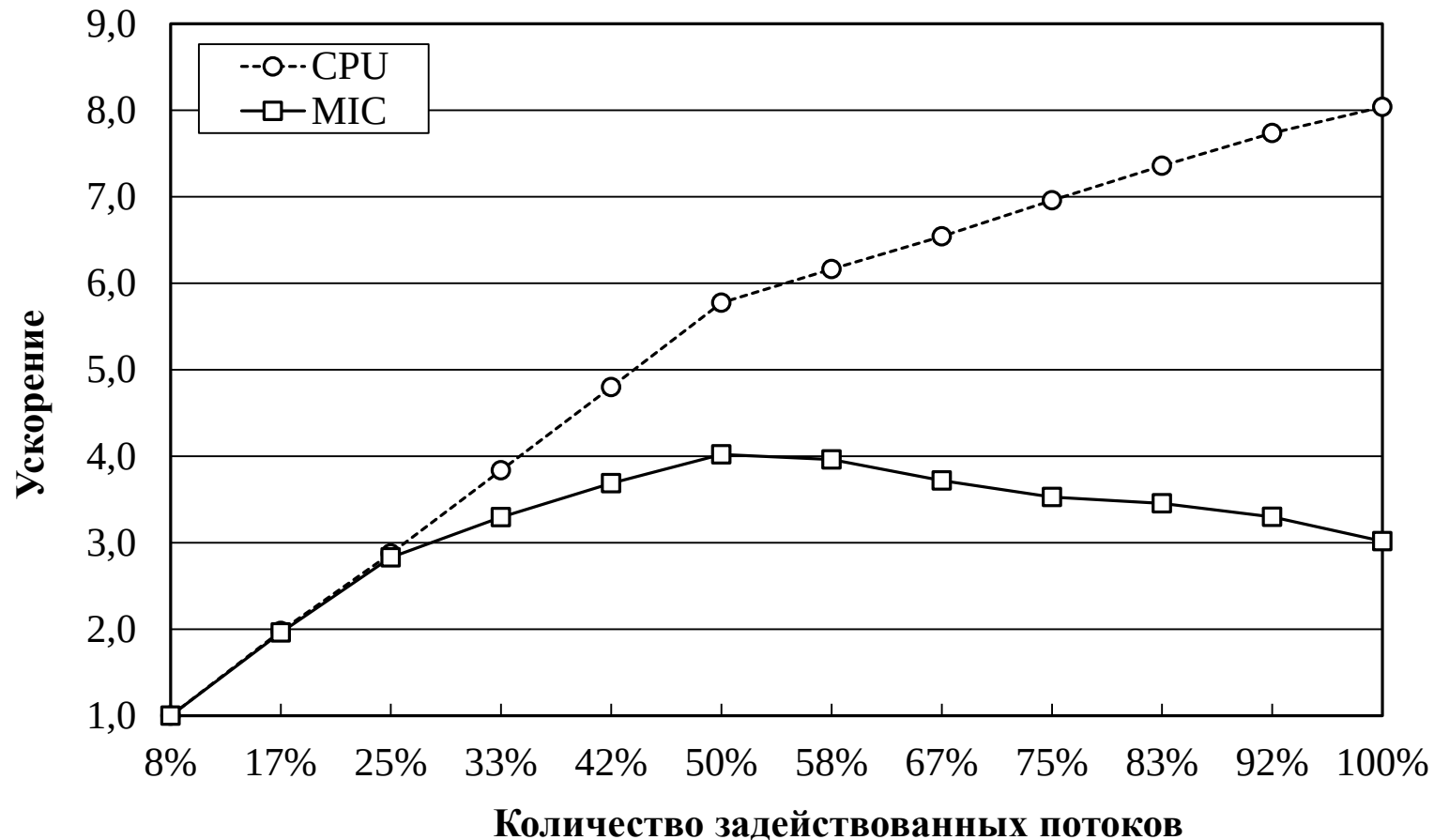
37 атрибутов

Балансировка загрузки процессорных ядер Xeon Phi



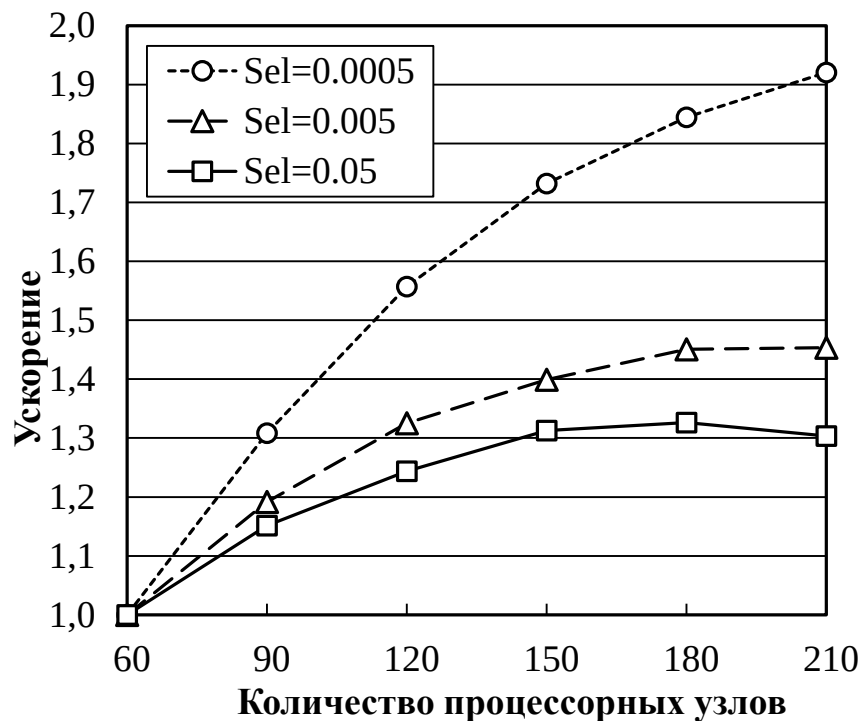
Размер хранилища: 63 млн. записей
Вычислительная система: «Торнадо ЮУрГУ»

Влияние гиперпоточности

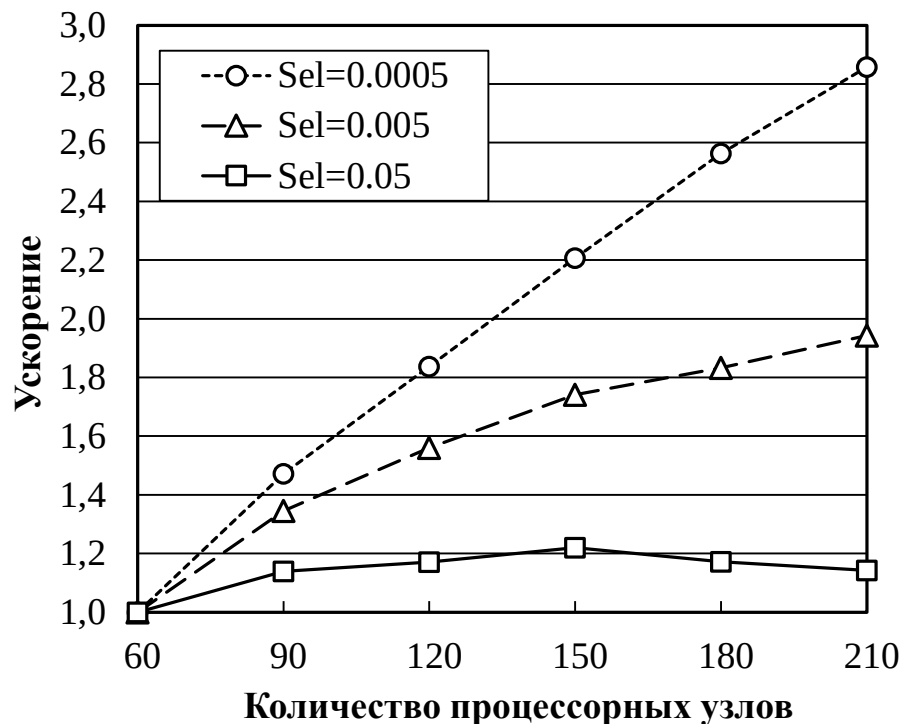


Размер хранилища: 63 млн. записей
Вычислительная система: «Торнадо ЮУрГУ»

Масштабируемость КСОП



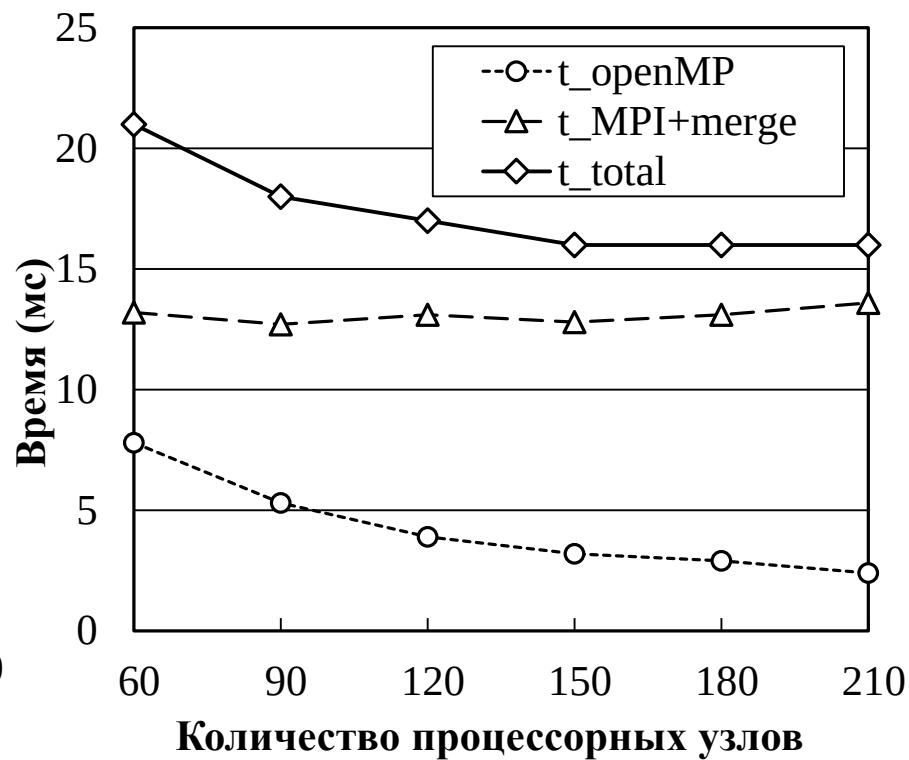
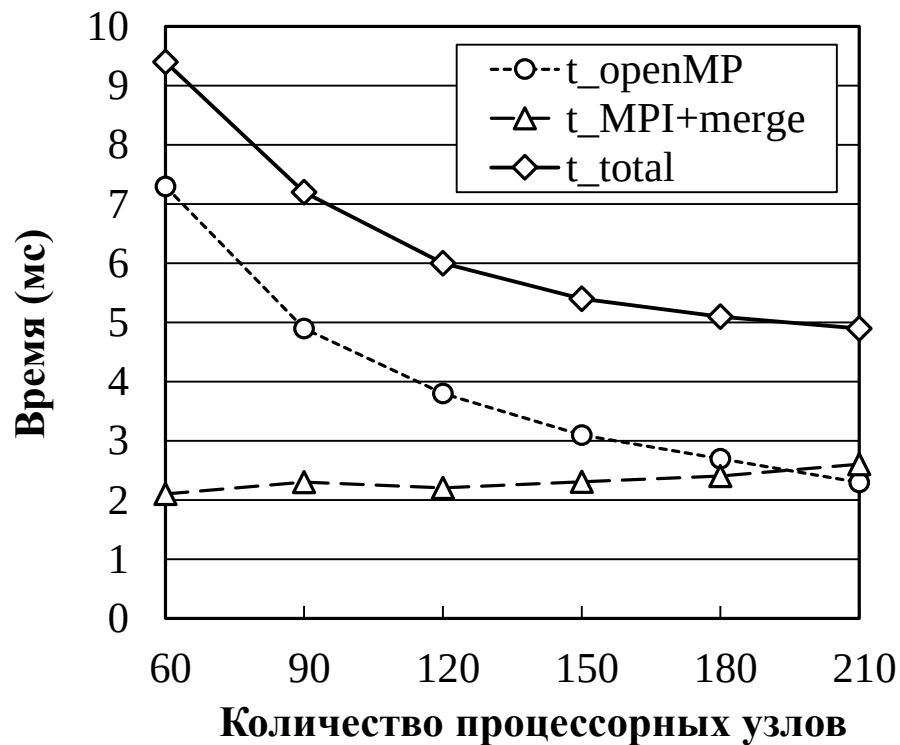
Размер хранилища:
63 млн. записей



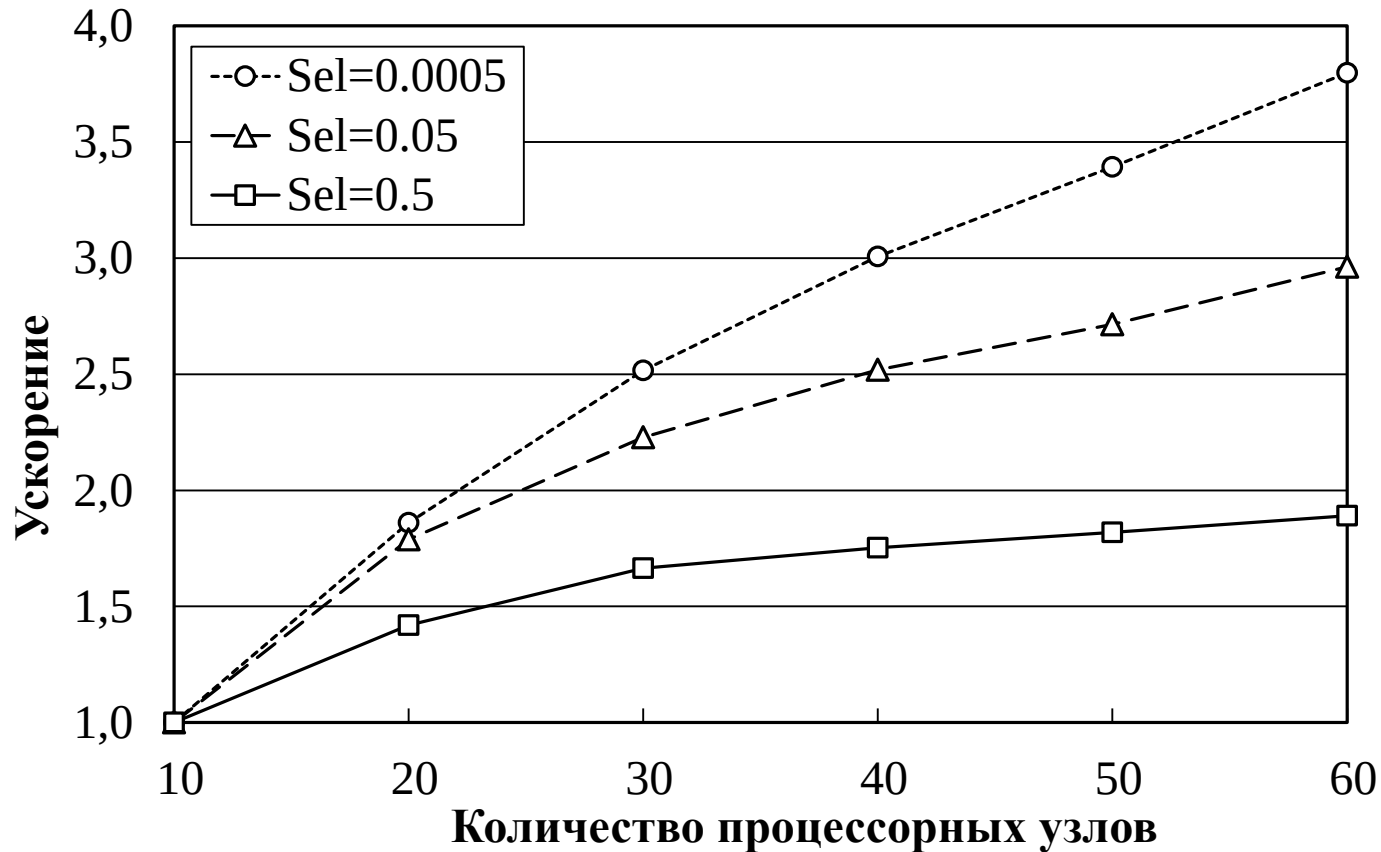
Размер хранилища:
630 млн. записей

Вычислительная система: «Торнадо ЮУрГУ»

Масштабируемость КСОП



Масштабируемость КСОП



Размер хранилища: 63 млн. записей

Вычислительная система: «RSC PetaStream»

Использование КСОП при выполнении SQL-запросов

Конфигурация	Время в минутах					
	Sel=0.0005		Sel=0.005		Sel=0.05	
	1-й запуск	2-й запуск	1-й запуск	2-й запуск	1-й запуск	2-й запуск
PostgreSQL	7.3	1.21	7.6	1.29	7.6	1.57
PostgreSQL & B-Trees	2.62	2.34	2.83	2.51	2.83	2.63
PostgreSQL & CCOP	0.073	0.008	0.65	0.05	2.03	1.72
Ускорение						
$\frac{t_{PostgreSQL}}{t_{PostgreSQL \& \ CCOP}}$	100	151	12	27	4	0.9
$\frac{t_{PostgreSQL \& \ B-Trees}}{t_{PostgreSQL \& \ CCOP}}$	36	293	4	50	1.4	1.53

Публикации

Статьи в журналах из перечня ВАК

1. *Иванова Е.В. Соколинский Л.Б.* Колоночный сопроцессор баз данных для кластерных вычислительных систем // Вестник Южно-Уральского государственного университета. Серия: Вычислительная математика и информатика. 2015. Т. 4, № 4. С. 25-49.
2. *Иванова Е.В. Соколинский Л.Б.* Использование сопроцессоров Intel Xeon Phi для выполнения естественного соединения над сжатыми данными // Вычислительные методы и программирование: Новые вычислительные технологии. 2015. Т. 16. Вып. 4. С. 534-542.
3. *Иванова Е.В. Соколинский Л.Б.* Параллельная декомпозиция реляционных операций на основе распределенных колоночных индексов // Вестник Южно-Уральского государственного университета. Серия: Вычислительная математика и информатика. 2015. Т. 4, № 4. С. 5-24.

Статья в издании, индексируемом в SCOPUS и Web of Science

4. *Ivanova E., Sokolinsky L.* Decomposition of Natural Join Based on Domain-Interval Fragmented Column Indices // Proceedings of the 38th International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO. IEEE, 2015. P. 223-226.

Статьи в изданиях, индексируемых в РИНЦ

1. *Иванова Е.В., Соколинский Л.Б.* Использование сопроцессоров Intel Xeon Phi для выполнения естественного соединения над сжатыми данными // Суперкомпьютерные дни в России: Труды международной конференции (28-29 сентября 2015 г., г. Москва). М.: Изд-во МГУ, 2015. С. 190-198.
2. *Иванова Е.В. Соколинский Л.Б.* Использование распределенных колоночных индексов для выполнения запросов к сверхбольшим базам данных // Параллельные вычислительные технологии (ПаВТ'2014): труды международной научной конференции (1–3 апреля 2014 г., г. Ростов-на-Дону). Челябинск: Издательский центр ЮУрГУ, 2014. С. 270–275.
3. *Иванова Е.В., Соколинский Л.Б.* Декомпозиция операции группировки на базе распределенных колоночных индексов // Наука ЮУрГУ. Челябинск: Издательский центр ЮУрГУ, 2015. С. 15-23.
4. *Иванова Е.В. Соколинский Л.Б.* Декомпозиция операций пересечения и соединения на основе доменно-интервальной фрагментации колоночных индексов // Вестник Южно-Уральского государственного университета. Серия: Вычислительная математика и информатика. 2015. Т. 4, № 1. С. 44-56.
5. *Иванова Е.В.* Исследование эффективности использования фрагментированных колоночных индексов при выполнении операции естественного соединения с использованием многоядерных ускорителей // Параллельные вычислительные технологии (ПаВТ'2015): труды международной научной конференции (30 марта - 3 апреля 2015 г., г. Екатеринбург). Челябинск: Издательский центр ЮУрГУ, 2015. С. 393–398.
6. *Иванова Е.В.* Использование распределенных колоночных хеш-индексов для обработки запросов к сверхбольшим базам данных // Научный сервис в сети Интернет: многообразие суперкомпьютерных миров: Труды Международной суперкомпьютерной конференции (22-27 сентября 2014 г., Новороссийск). М.: Изд-во МГУ, 2014. С. 102-104.

Основные результаты, выносимые на защиту

- 1) Разработана доменно-колоночная модель представления данных, на базе которой выполнена декомпозиция основных реляционных операций с помощью распределенных колоночных индексов.
- 2) Разработаны высокомасштабируемые параллельные алгоритмы выполнения основных реляционных операций, использующие распределенные колоночные индексы.
- 3) Выполнена реализация колоночного сопроцессора для кластерных вычислительных систем. Общий объем кода на языке Си составил около 2500 строк. Исходные тексты прототипа свободно доступны в Интернет по адресу: <https://github.com/elena-ivanova/columnindices/>.
- 4) Проведены вычислительные эксперименты, подтверждающие эффективность предложенных подходов.