

# Построение отказоустойчивых распределенных систем хранения данных на основе модулярной арифметики

**Назаров Антон Сергеевич**

к.т.н., м.н.с. УНЦ «Вычислительная математика и  
параллельное программирование на супер ЭВМ»

ФГАОУ ВО «Северо-Кавказский федеральный университет»

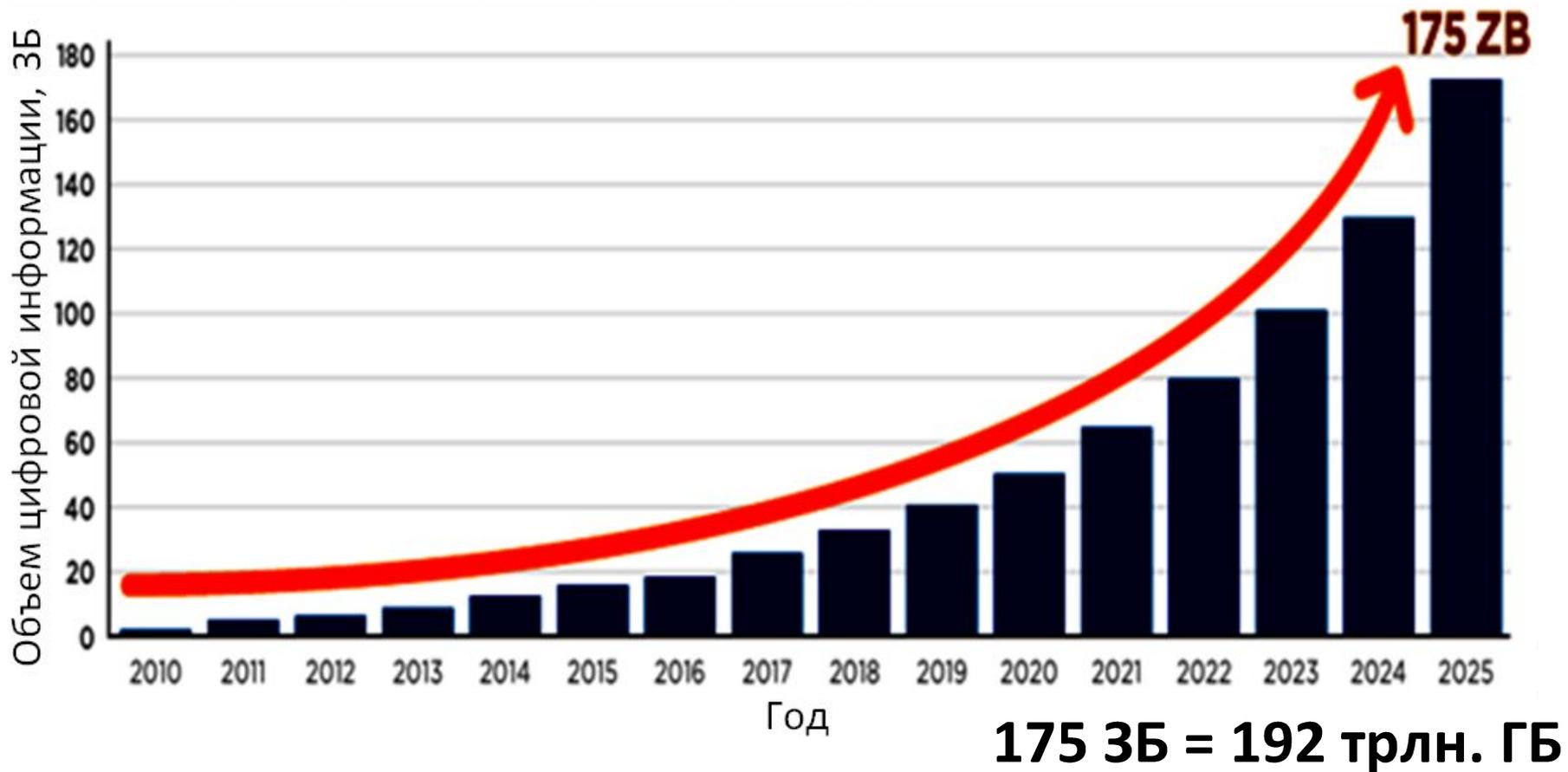
Научный руководитель: заслуженный деятель науки и техники РФ,

доктор технических наук, профессор

**Червяков Николай Иванович**

# Актуальность

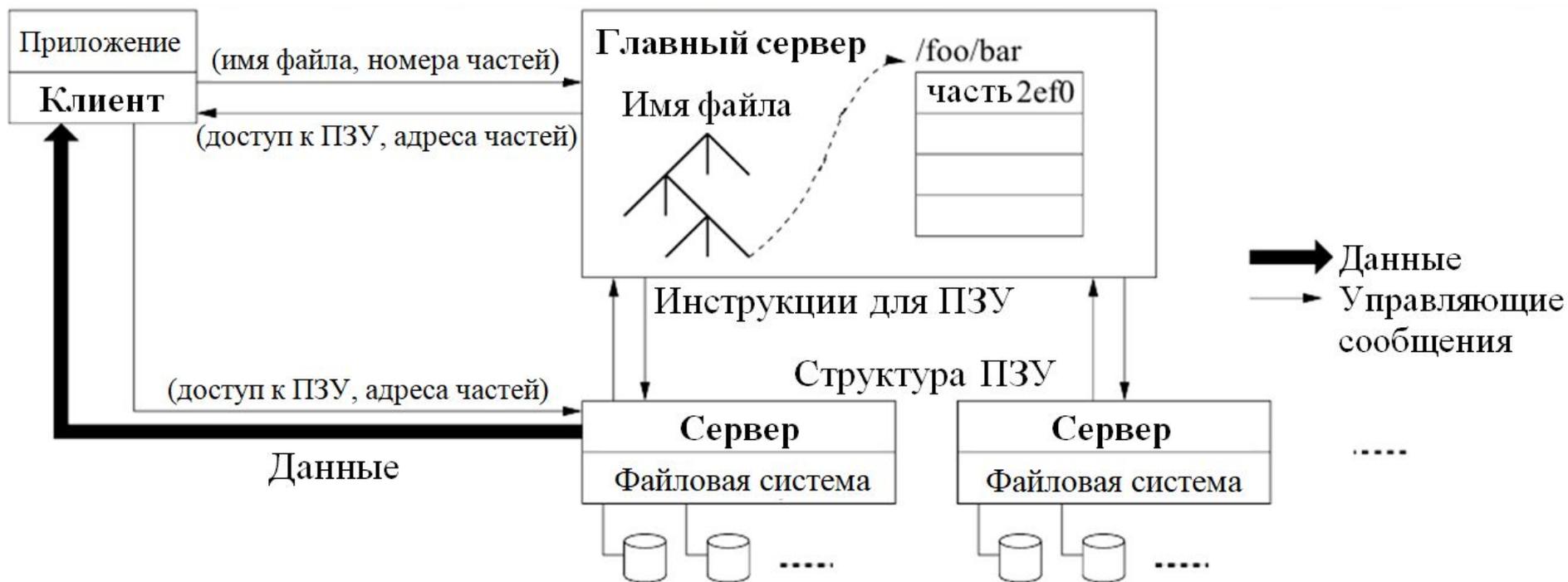
## РОСТ МИРОВОГО ОБЪЕМА ЦИФРОВОЙ ИНФОРМАЦИИ



\*по данным International Data Corporation (IDC) на ноябрь 2018

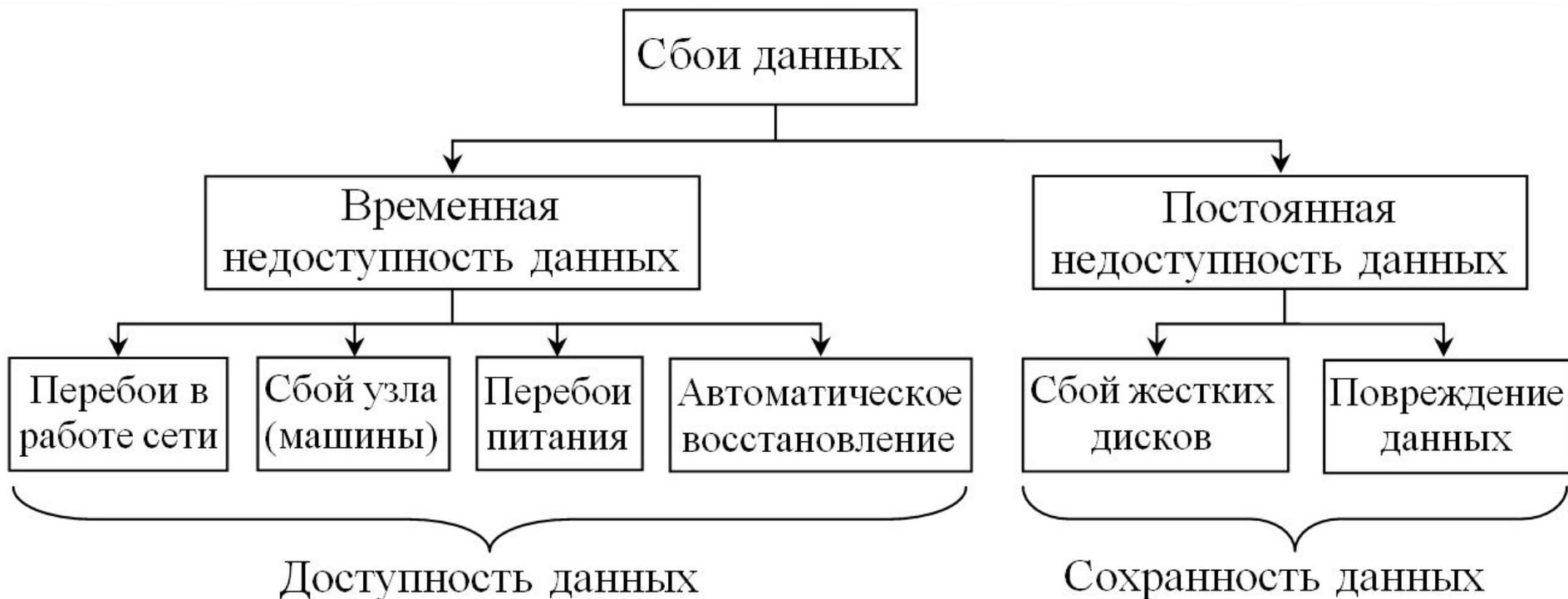
# Актуальность

## СТРУКТУРА РАСПРЕДЕЛЕННЫХ СИСТЕМ ХРАНЕНИЯ ДАННЫХ



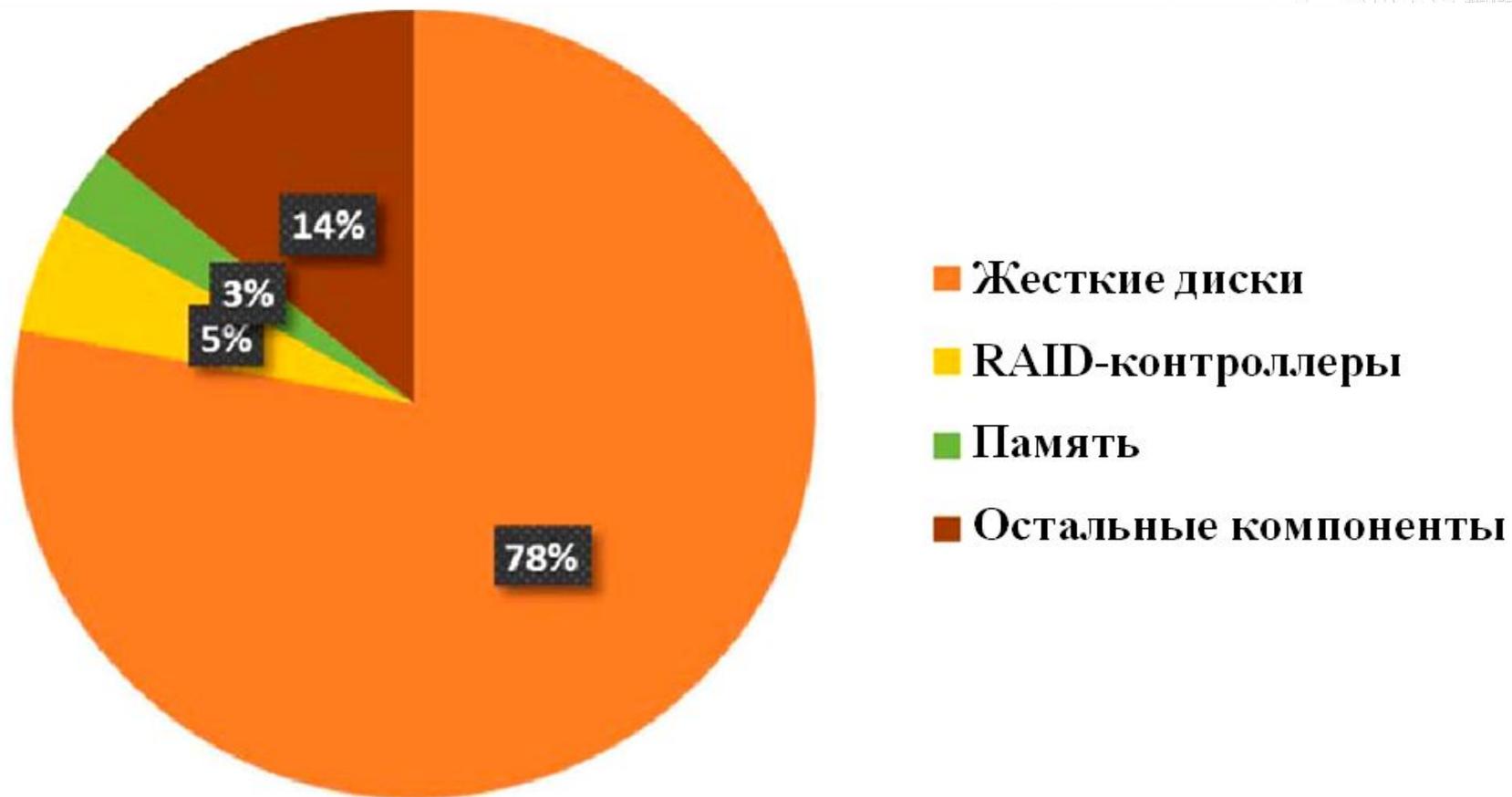
\*Ghemawat, S., Gobioff, H., and Leung, S.-T. The Google File System. In 19th Symposium on Operating Systems Principles, Lake George, NY, pp. 29-43, 2003.

## КЛАССИФИКАЦИЯ СБОЕВ РАСПРЕДЕЛЕННЫХ СИСТЕМ ХРАНЕНИЯ ДАННЫХ



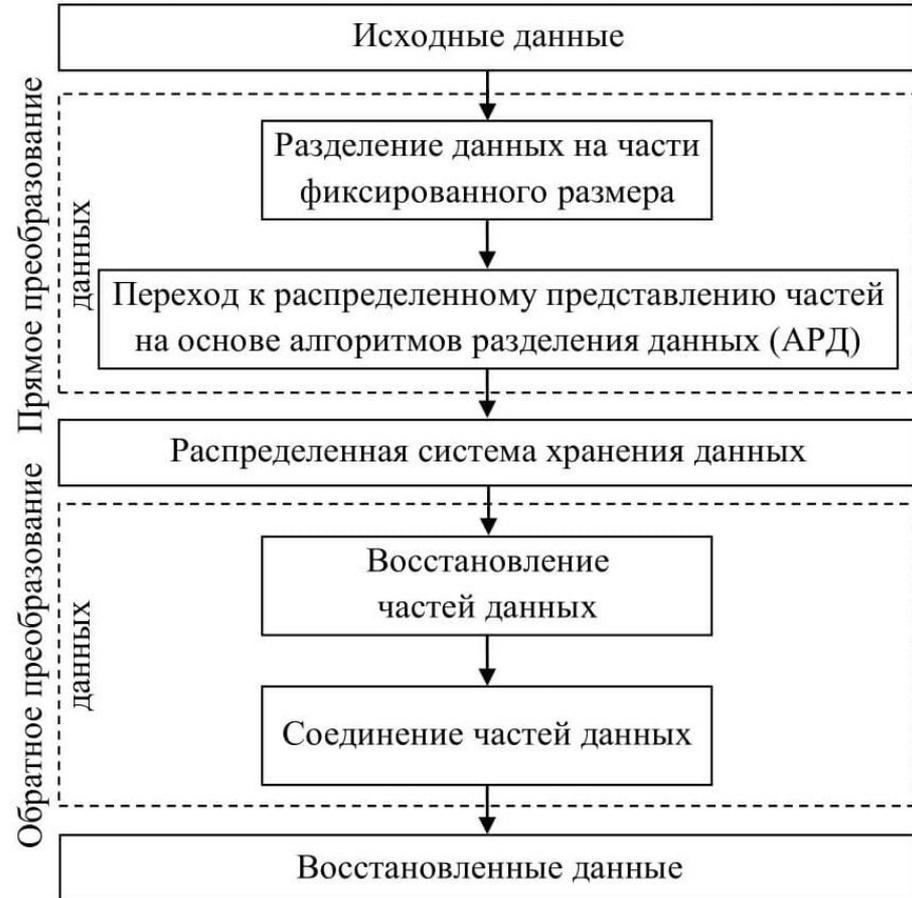
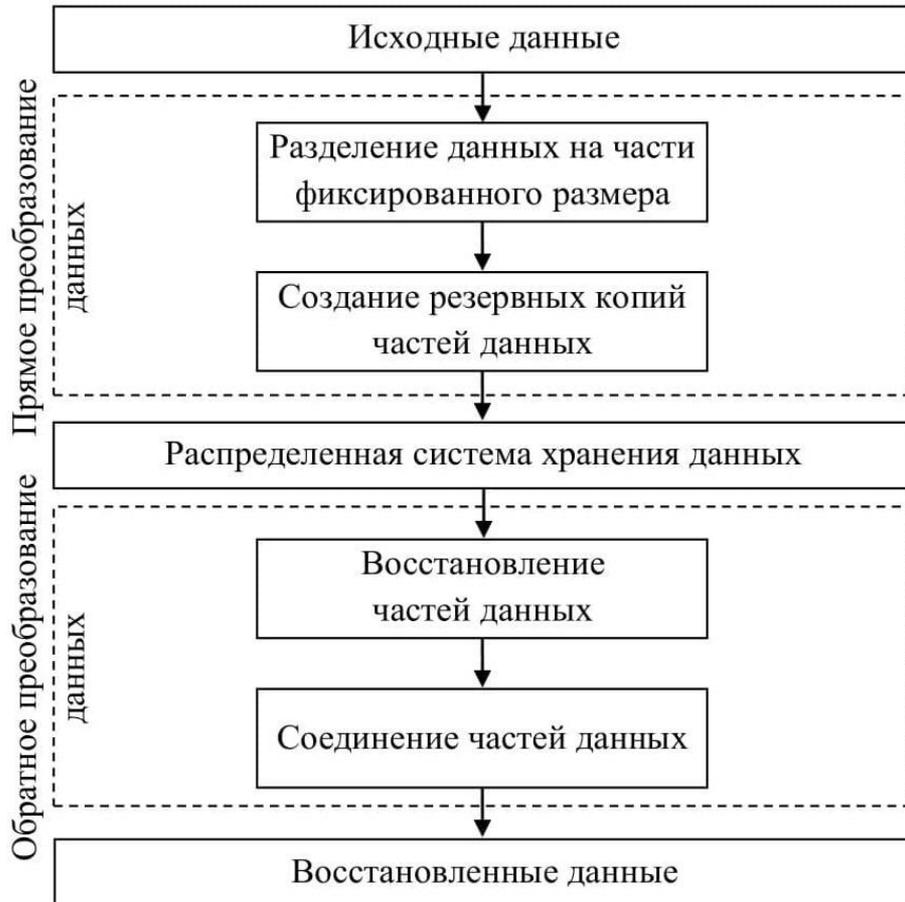
\* Nachiappan, R. Cloud storage reliability for Big Data applications: A state of the art survey / R. Nachiappan, B. Javadi, R.N. Calheiros [et al.] // Journal of Network and Computer Applications. – 2017. – Vol. 97. – P. 35-47.

# ПРИЧИНЫ СБОЕВ В РАСПРЕДЕЛЕННЫХ СИСТЕМАХ ХРАНЕНИЯ ДАННЫХ

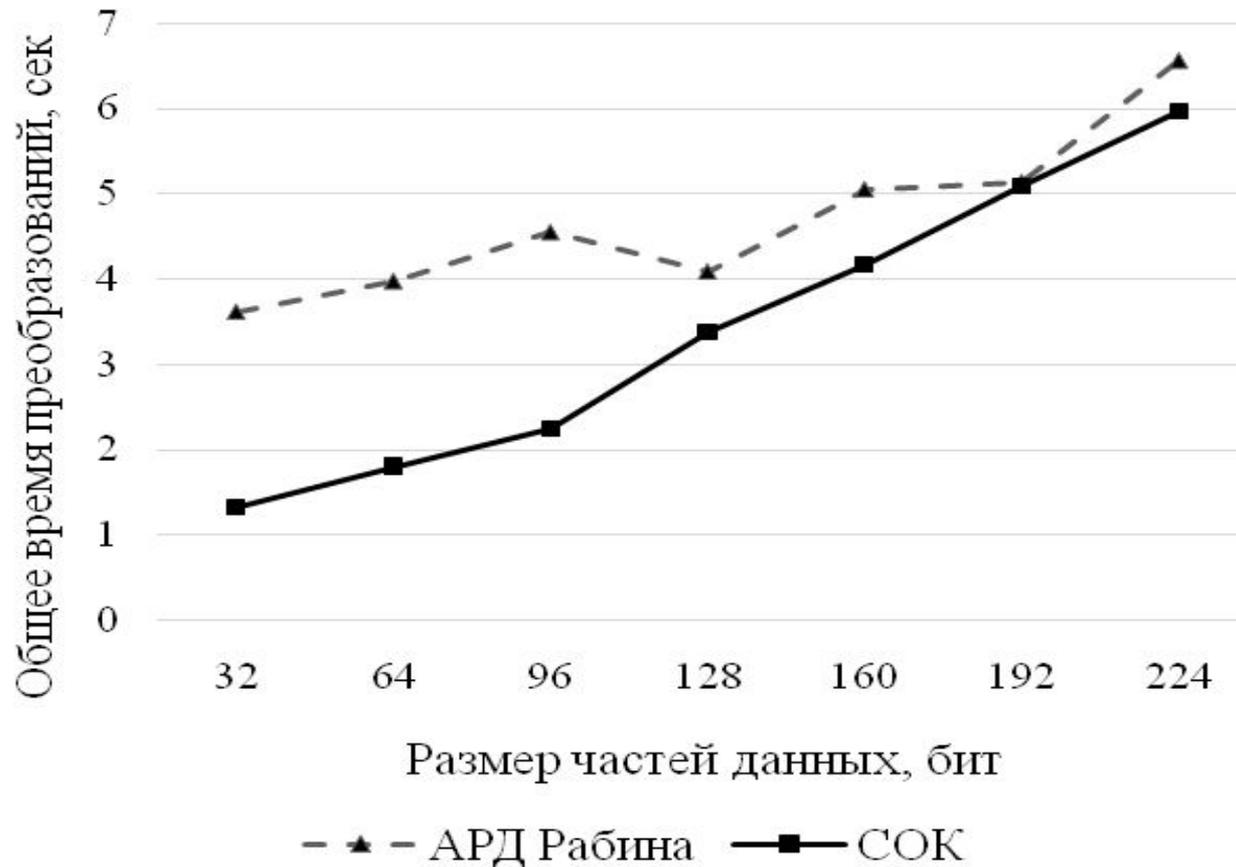


\* Sharma, Y. Reliability and energy efficiency in cloud computing systems: Survey and taxonomy / Y. Sharma, B. Javadi, W. Si [et al.] // Journal of Network and Computer Applications. – 2016. – Vol. 74. – P. 66-85.

# СТАНДАРТНЫЙ И ПРЕДЛАГАЕМЫЙ ПОДХОДЫ К ОБЕСПЕЧЕНИЮ НАДЕЖНОСТИ РАСПРЕДЕЛЕННЫХ СИСТЕМ ХРАНЕНИЯ ДАННЫХ



# ПРОИЗВОДИТЕЛЬНОСТЬ ОСНОВНЫХ АЛГОРИТМОВ РАЗДЕЛЕНИЯ ДАННЫХ



\* Deryabin, M. Comparative Performance Analysis of Information Dispersal Methods / M. Deryabin, N. Chervyakov, A. Nazarov [et al.] // FRUCT: Proceedings of the 24th Conference of Open Innovations Association. – Moscow, Russia: IEEE, 2019. – P. 67-74.

# СИСТЕМА ОСТАТОЧНЫХ КЛАССОВ



Система Остаточных Классов (СОК)\*:

- Непозиционная система счисления
- Числа представлены наборами остатков от деления на основания СОК
- Возможность распараллеливания арифметических вычислений
- Возможность контроля целостности
- Основные приложения СОК:
  - Цифровая обработка сигналов
  - Обработка изображений
  - Криптография
  - **Повышение надежности**
  - Облачные вычисления
  - Big Data
  - и т.д.

$m_i$  – основания СОК

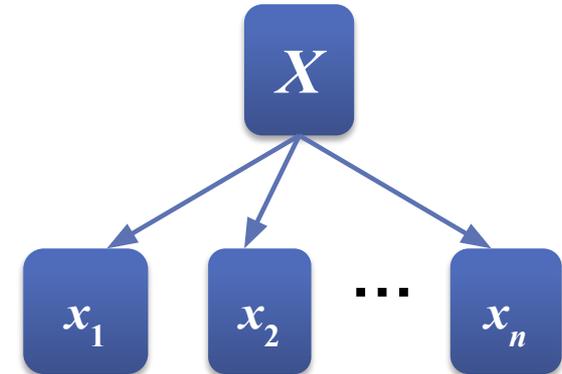
(попарно взаимно простые

числа,  $\text{НОД}(m_i, m_j) = 1, \forall i \neq j$ )

$X$  – число в позиционной системе счисления ( $0 \leq X < m_1 \cdot m_2 \cdot \dots \cdot m_n$ )

$x_i = X \bmod m_i$  – компоненты числа в СОК (цифры)

$$X = (x_1, x_2, \dots, x_n)$$



\* Omondi, A. Residue Number Systems. Theory and Implementation / A. Omondi, B. Premkumar. – London, England: Imperial College Press, 2007. – 296 p.

# ИЗБЫТОЧНАЯ СИСТЕМА ОСТАТОЧНЫХ КЛАССОВ

$m_i$  – основания избыточной СОК (ИСОК)

(попарно взаимно простые числа,  $\text{НОД}(m_i, m_j) = 1, \forall i \neq j$ )

$M_k = m_1 \cdot m_2 \cdot \dots \cdot m_k$  – рабочий диапазон ИСОК

$M_n = M_k \cdot m_{k+1} \cdot m_{k+2} \cdot \dots \cdot m_n$  – полный диапазон ИСОК

$n$  – общее кол-во оснований,  $k$  – кол-во рабочих оснований

$X$  – число в позиционной системе счисления ( $0 \leq X < m_1 \cdot m_2 \cdot \dots \cdot m_k$ )

$x_i = X \bmod m_i$  – компоненты числа в ИСОК (цифры)

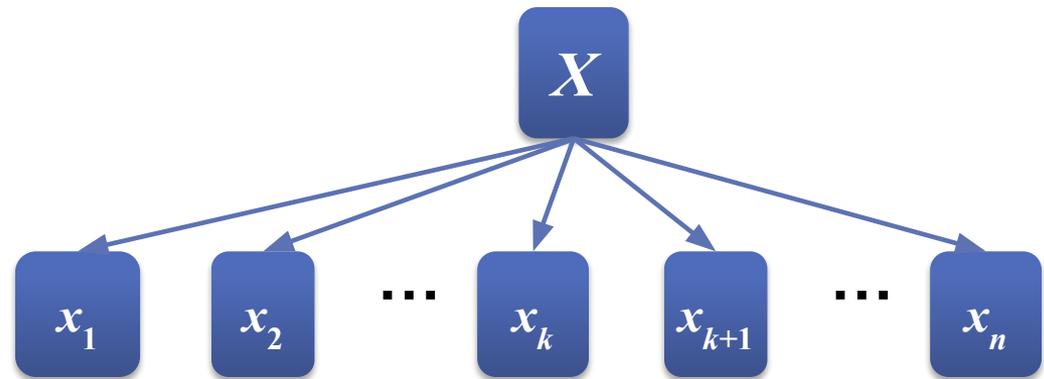
$$X = (x_1, x_2, \dots, x_k, x_{k+1}, \dots, x_n)$$

**Корректирующая  
способность  $(k,n)$ -ИСОК\*:**

**Обнаружение ошибки:  $r$**

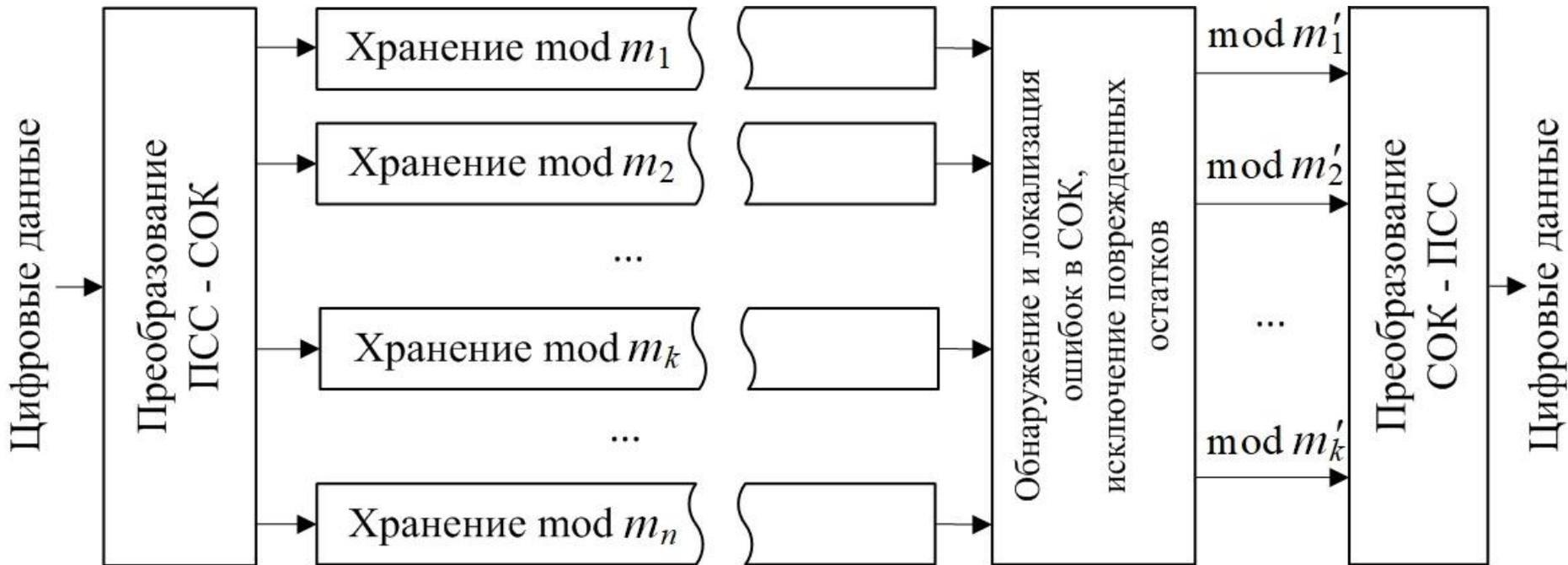
**Исправление ошибки:  $\lfloor r/2 \rfloor$ ,**

**где  $r = n - k$**



\* Ding, C. Chinese remainder theorem: applications in computing, coding, cryptography / C. Ding, D. Pei, A. Salomaa. – Singapore: World Scientific, 1996. – 214 p.

# МОДЕЛЬ РАСПРЕДЕЛЕННОГО ХРАНЕНИЯ ДАННЫХ НА ОСНОВЕ ИЗБЫТОЧНОЙ СИСТЕМЫ ОСТАТОЧНЫХ КЛАССОВ

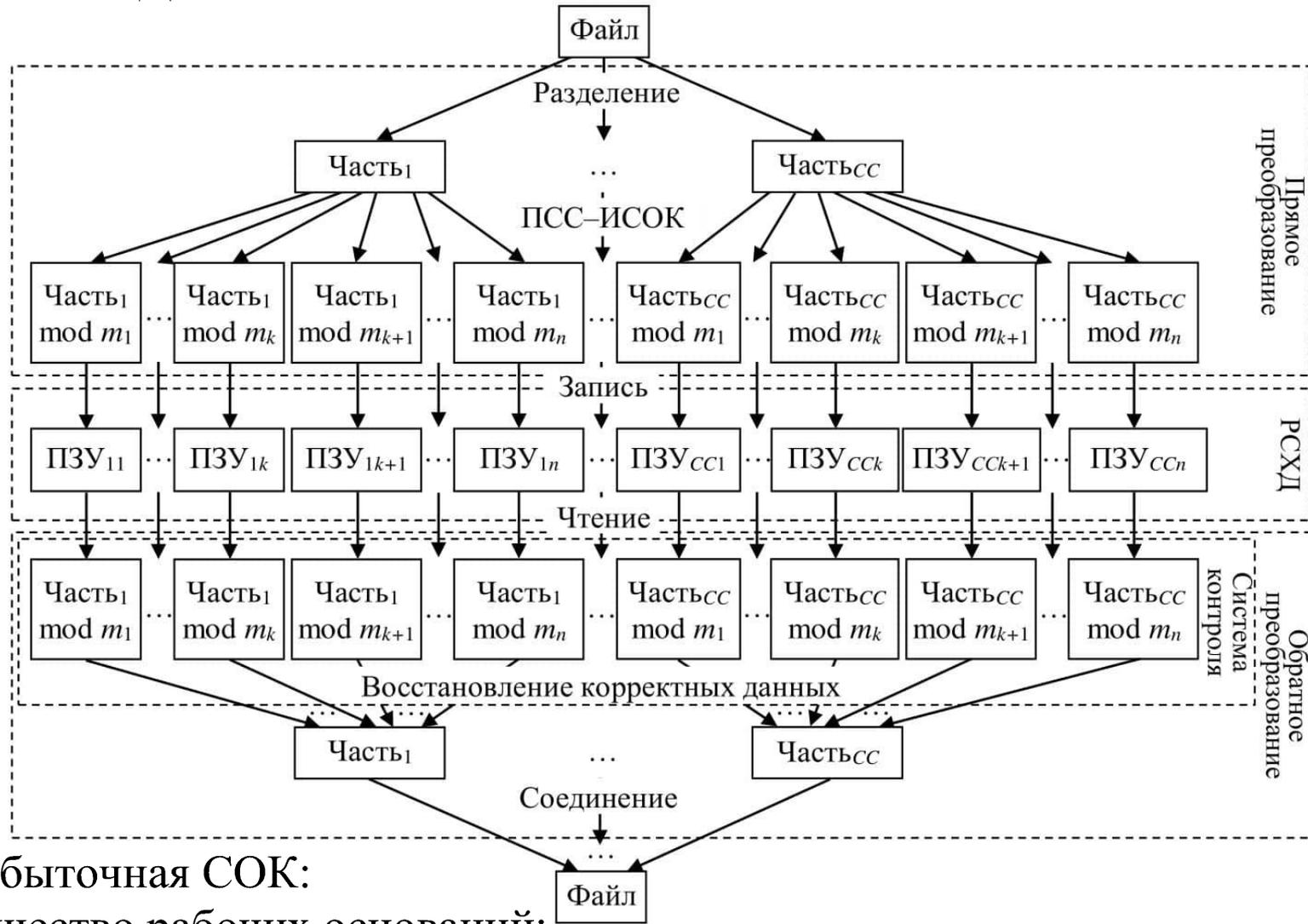


$(k, n)$ -избыточная СОК:

$k$  – количество рабочих оснований;

$n$  – общее количество оснований.

# ОБОБЩЕННАЯ СХЕМА РАСПРЕДЕЛЕННОГО ХРАНЕНИЯ ДАННЫХ НА ОСНОВЕ ИЗБЫТОЧНОЙ СОК



$(k, n)$ -избыточная СОК:

$k$  – количество рабочих оснований;

$n$  – общее количество оснований;

$СС$  – кол-во частей после разделения файла.

# КЛАССИФИКАЦИЯ МЕТОДОВ ИСПРАВЛЕНИЯ ОШИБОК НА ОСНОВЕ ИСОК



## **Метод Непрерывных Дробей:**

- Goldreich O., Ron D., Sudan M. Chinese remaindering with errors // IEEE Transactions on Information Theory. 2000. Vol. 46, no. 4. P. 1330-1338.
- Mandelbaum D. On a class of arithmetic codes and a decoding algorithm // IEEE Transactions on Information Theory. 1976. Vol. 22, no. 1. P. 85-88.

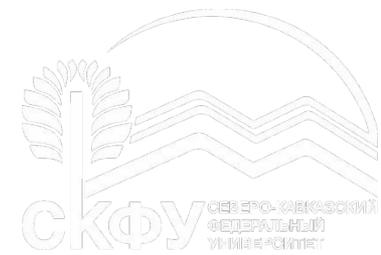
## **Синдромное Декодирование:**

- Tay T.F., Chang C.H. A Non-Iterative Multiple Residue Digit Error Detection and Correction Algorithm in RRNS // IEEE Transactions on Computers. 2016. Vol. 65, no. 2. P. 396-408.
- Червяков Н.И., Сахнюк П.А., Шапошников А.В., Макоха А.Н. Нейро-компьютеры в остаточных классах. – М.: «Радиотехника», 2003. – 272 с.

## **Метод Модулярных Проекций:**

- Goh V.T., Siddiqi M.U. Multiple error detection and correction based on redundant residue number systems // IEEE Transactions on Communications. 2008. Vol. 56, no. 3. P. 325-330.
- Chervyakov, N.I. The architecture of a fault-tolerant modular neurocomputer based on modular number projections / N.I. Chervyakov, P.A. Lyakhov, A.S.Nazarov [et al.] // Neurocomputing. – 2018. – Vol. 272. – P. 96-107.

# МЕТОД НЕПРЕРЫВНЫХ ДРОБЕЙ



**Достоинства:** –

**Недостатки:**

- возрастание количества итераций при увеличении кратности ошибок и разрядности остатков (!!! критический недостаток);
- неэффективность при аппаратной реализации.



## *Синдромное декодирование с ограничением на надежность избыточных остатков*

### **Достоинства:**

- высокая скорость обнаружения, локализации и исправления ошибок;
- уменьшение объема коррекционных таблиц по сравнению с полным синдромным декодированием;

### **Недостатки:**

- требование абсолютной надежности избыточных остатков (!!! критический недостаток);
- произведение контрольных модулей должно более чем вдвое превышать величину рабочего диапазона:  $2M_k < m_{k+1}m_{k+2}\dots m_n$ ;
- быстрое увеличение объема коррекционных таблиц с увеличением кратности ошибки.

## *Полное синдромное декодирование*

### **Достоинства:**

- высокая скорость обнаружения, локализации и исправления ошибок;
- отсутствие ограничения, связанного с абсолютной надежностью контрольных остатков;

### **Недостатки:**

- произведение контрольных модулей должно более чем вдвое превышать величину рабочего диапазона:  $2M_k < m_{k+1}m_{k+2}\dots m_n$ ;
- быстрое увеличение объема коррекционных таблиц с увеличением кратности ошибки (!!! критический недостаток).

## *Классический метод проекций*

### **Достоинства:**

- высокая скорость обнаружения, локализации и исправления ошибок;
- менее жесткие по сравнению с синдромным декодированием ограничения на выбор избыточных оснований:  $m_{k+1}m_{k+2}\dots m_n > m_i$ ,  
 $\square i \leq k$ ;

### **Недостатки:**

- быстрое увеличение количества проекций с увеличением кратности ошибки (!!! критический недостаток);
- увеличение объема предвычисленных констант с увеличением количества проекций.

## *Метод проекций с максимальным правдоподобием*

### **Достоинства:**

- высокая скорость обнаружения, локализации и исправления ошибок;
- менее жесткие по сравнению с синдромным декодированием ограничения на выбор избыточных оснований:  $m_{k+1}m_{k+2}\dots m_n > m_i$ ,  
 $\square i \leq k$ ;
- уменьшение количества проекций по сравнению с классическим методом проекций.

### **Недостатки:**

- необходимость дополнительного шага расчета расстояний Хэмминга между ошибочным числом и каждой из проекций;
- увеличение объема предвычисленных констант с увеличением количества проекций.

# МЕТОДЫ ОБНАРУЖЕНИЯ, ЛОКАЛИЗАЦИИ И ИСПРАВЛЕНИЯ ОШИБОК НА ОСНОВЕ ИЗБЫТОЧНОЙ СОК В РАСПРЕДЕЛЕННЫХ СИСТЕМАХ ХРАНЕНИЯ ДАННЫХ



Метод проекций $(k, n): C_n^t$	Метод проекций с максимальным правдоподобием $(k, n): \lfloor n/k \rfloor^*$
(2,6)-избыточная СОК (2, 3, 5, 7, 11, 13); Максимальная кратность исправляемой ошибки: $t = \lfloor (n - k)/2 \rfloor = 2$ .	
$(-, -, 5, 7, 11, 13), (-, 3, -, 7, 11, 13),$ $(-, 3, 5, -, 11, 13), (-, 3, 5, 7, -, 13),$ $(-, 3, 5, 7, 11, -), (2, -, -, 7, 11, 13),$ $(2, -, 5, -, 11, 13), (2, -, 5, 7, -, 13),$ $(2, -, 5, 7, 11, -), (2, 3, -, -, 11, 13),$ $(2, 3, -, 7, -, 13), (2, 3, -, 7, 11, -),$ $(2, 3, 5, -, -, 13), (2, 3, 5, -, 11, -),$ $(2, 3, 5, 7, -, -).$	$(2, 3, -, -, -, -),$ $(-, -, 5, 7, -, -),$ $(-, -, -, -, 11, 13).$ <p style="text-align: center;">+</p> <p style="text-align: center;">Расстояние Хэмминга</p>
<b>Количество проекций</b>	
<b>15</b>	<b>3</b>

\* Оценка справедлива для (2,6)-ИСОК, для других  $(k, n)$ -ИСОК может отличаться. Подробнее в работе: Goh, V.T. Multiple error detection and correction based on redundant residue number systems / V.T. Goh, M.U. Siddiqi // IEEE Transactions on Communications. – 2008. – Vol. 56. – No. 3. – P. 325-330.

# РАСПРЕДЕЛЕННАЯ СИСТЕМА ХРАНЕНИЯ ДАННЫХ, ОСНОВАННАЯ НА РЕПЛИКАЦИИ



Вероятность отказа\* при запросе PFD равна:

$$PFD = 1 - \left( \sum_{j=1}^{RF} C_{RF}^j \cdot (1 - AFR)^j AFR^{RF-j} \cdot \sum_{i=\lfloor \frac{j}{2} \rfloor + 1}^j C_j^i (1 - er)^i er^{j-i} \right)^{CC},$$

где  $AFR^{**}$  – вероятность выхода из строя одного жесткого диска в течение времени использования равного  $T_0$ ,

$$AFR = 1 - e^{-T_0 / MTBF},$$

где MTBF (Mean Time Between Failures) – это среднее время между отказами,  $RF$  – фактор репликации,  $CC$  (Chunks Count) – количество частей, получившееся после разрезания файла,  $er$  – вероятность искажения данных.

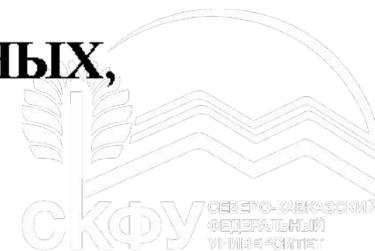
Избыточность данных:  $Redundancy = (RF - 1) \cdot 100\%$ .

\* Назаров А.С. Вероятностный подход к оценке отказоустойчивости различных моделей распределенного хранения данных / А.С. Назаров, М.А. Дерябин, М.Г. Бабенко [и др.] // Инженерный вестник Дона. – 2019. – № 8(59). – С. 19:1-30.

\*\* Szabados, D. Diving into “MTBF” and “AFR”: Storage Reliability Specs Explained [Электронный ресурс] / D. Szabados // Inside IT Storage. Seagate Enterprise – 2010. – Режим доступа:

<https://web.archive.org/web/20100501151901/http://enterprise.media.seagate.com/2010/04/inside-it-storage/diving-into-mtbf-and-afr-storage-reliability-specs-explained/> – (Дата обращения: 15.06.2019).

# РАСПРЕДЕЛЕННАЯ СИСТЕМА ХРАНЕНИЯ ДАННЫХ, ОСНОВАННАЯ НА ИЗБЫТОЧНОЙ СОК



Вероятность отказа\* при запросе PFD\*\* равна:

$$PFD = 1 - \left( \sum_{j=k}^n \left( C_n^j \cdot (1 - AFR)^j AFR^{n-j} \cdot \sum_{i=j - \lfloor \frac{j-k}{2} \rfloor}^j C_j^i (1 - er)^i er^{j-i} \right) \right)^{CC}.$$

где AFR – вероятность выхода из строя одного жесткого диска в течение времени использования равного  $T_0$ ,  $n$  – общее количество подчастей, формируемое для каждой части,  $k$  – количество подчастей, достаточное для восстановления части,  $CC$  (Chunks Count) – количество частей, получившееся после разрезания файла,  $er$  – вероятность искажения данных.

$$\text{Избыточность данных: Redundancy} \approx \left( \frac{n}{k} - 1 \right) \cdot 100\%.$$

\* Назаров А.С. Вероятностный подход к оценке отказоустойчивости различных моделей распределенного хранения данных / А.С. Назаров, М.А. Дерябин, М.Г. Бабенко [и др.] // Инженерный вестник Дона. – 2019. – № 8(59). – С. 19:1-30.

\*\* Braband J. Probability of failure on demand - The why and the how. / J. Braband, R. VomHövel, H. Schäbe // Lecture Notes in Computer Science. Springer. – 2009. – Vol. 5775. – P. 46–54.

# СРАВНЕНИЕ ОТКАЗОУСТОЙЧИВОСТИ МОДЕЛЕЙ НА ОСНОВЕ РЕПЛИКАЦИИ И ИЗБЫТОЧНОЙ СОК

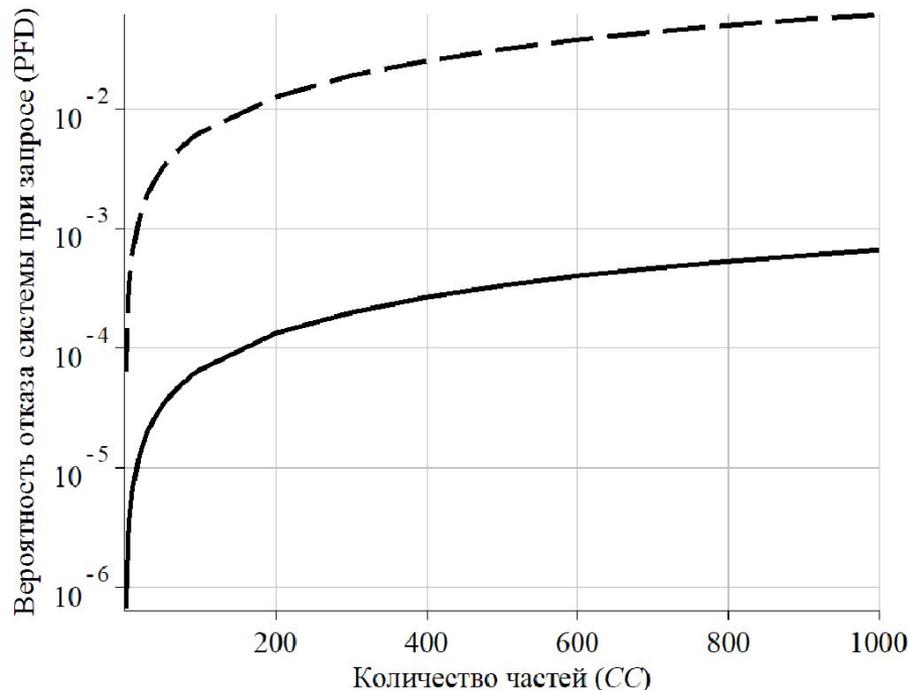
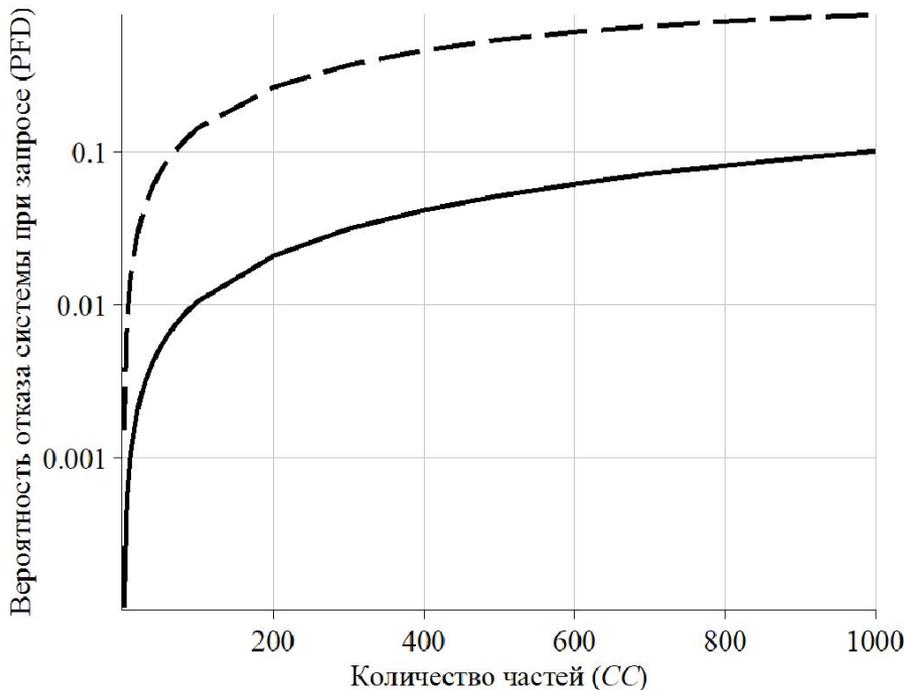


Период непрерывного функционирования СХД:  $T_0 = 8766$  ч. (1 год).

Вероятность искажения данных:  $er = 0.001$

Использовались:

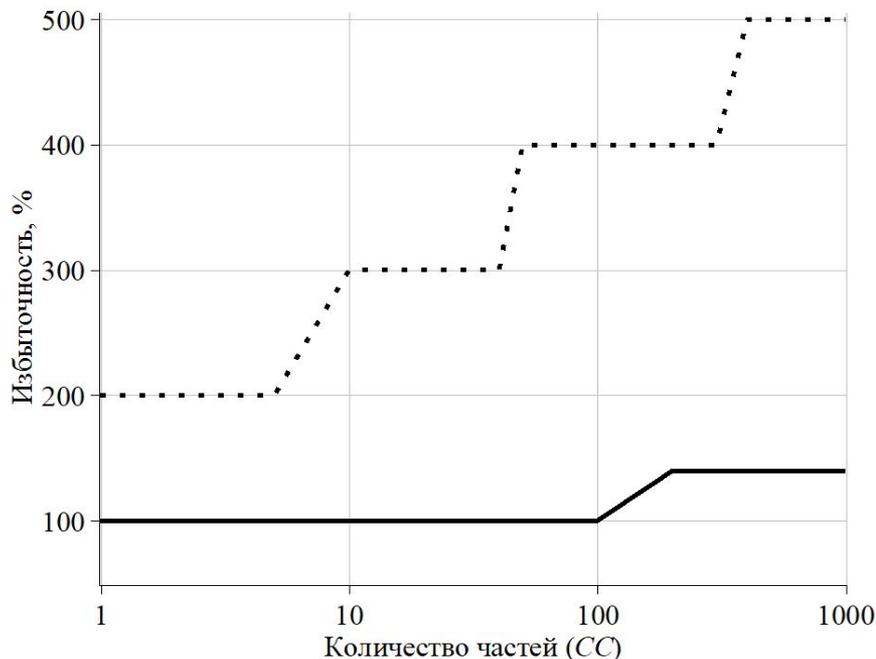
(2,6)-избыточная СОК (избыточность 200%) и  $RF = 3$  – резервирование (избыточность 200%)



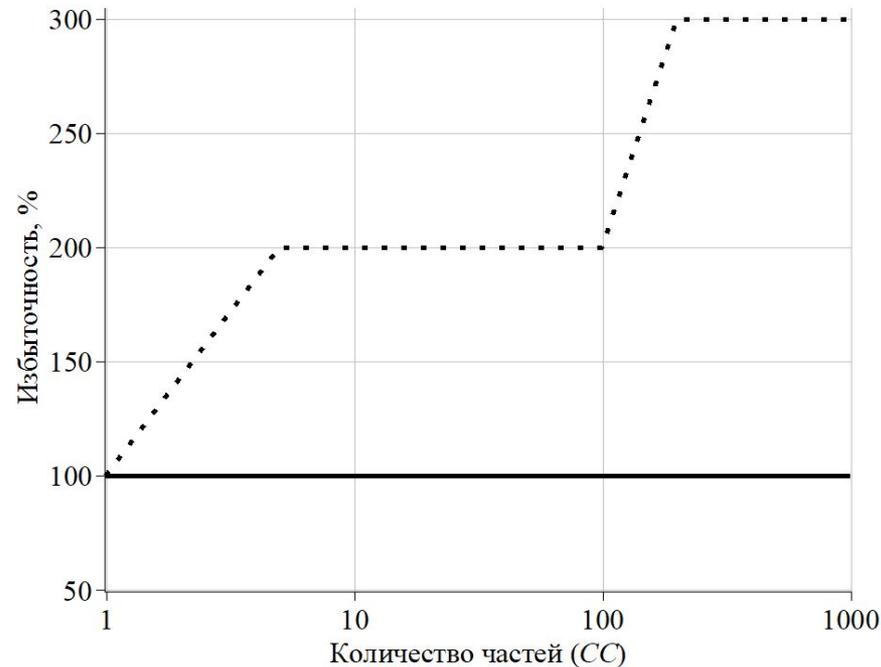
--- Стандартная модель распределенного хранения на основе тройного резервирования, AFR = 0.1  
 — Предложенная модель распределенного хранения на основе (2,6)-избыточной СОК, AFR = 0.1

--- Стандартная модель распределенного хранения на основе тройного резервирования, AFR = 0.01  
 — Предложенная модель распределенного хранения на основе (2,6)-избыточной СОК, AFR = 0.01

# СРАВНЕНИЕ МОДЕЛЕЙ РАСПРЕДЕЛЕННОГО ХРАНЕНИЯ ДАННЫХ НА ОСНОВЕ РЕПЛИКАЦИИ И ИЗБЫТОЧНОЙ СОК



• • Стандартная модель с резервированием, AFR = 0.1  
— Предложенная модель с разделением данных, AFR = 0.1



• • Стандартная модель с резервированием, AFR = 0.01  
— Предложенная модель с разделением данных, AFR = 0.01

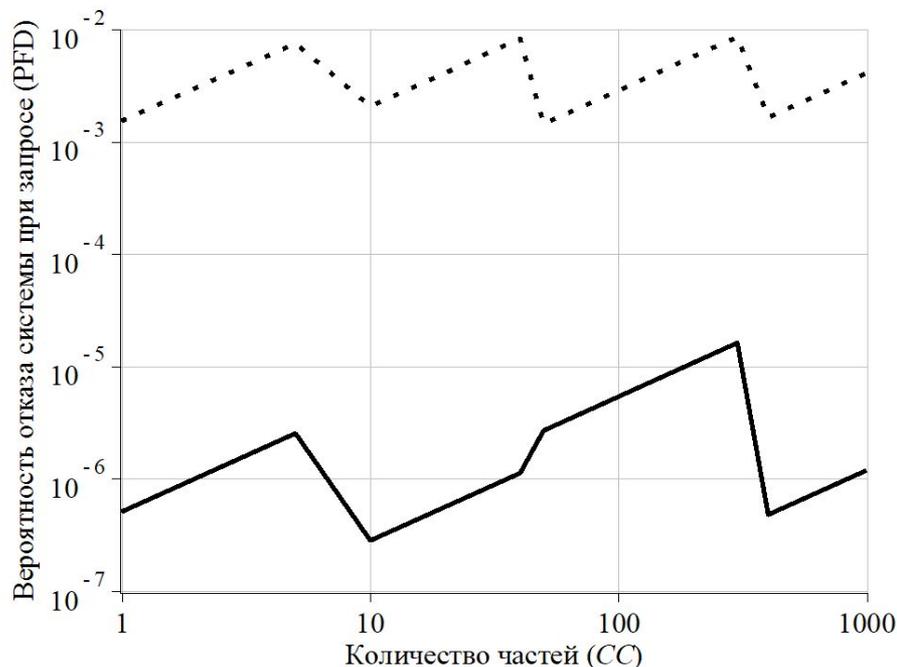
Условия:  $PFD \leq 10^{-2}$

**Преимущество ИСОК по сравнению с резервированием:**

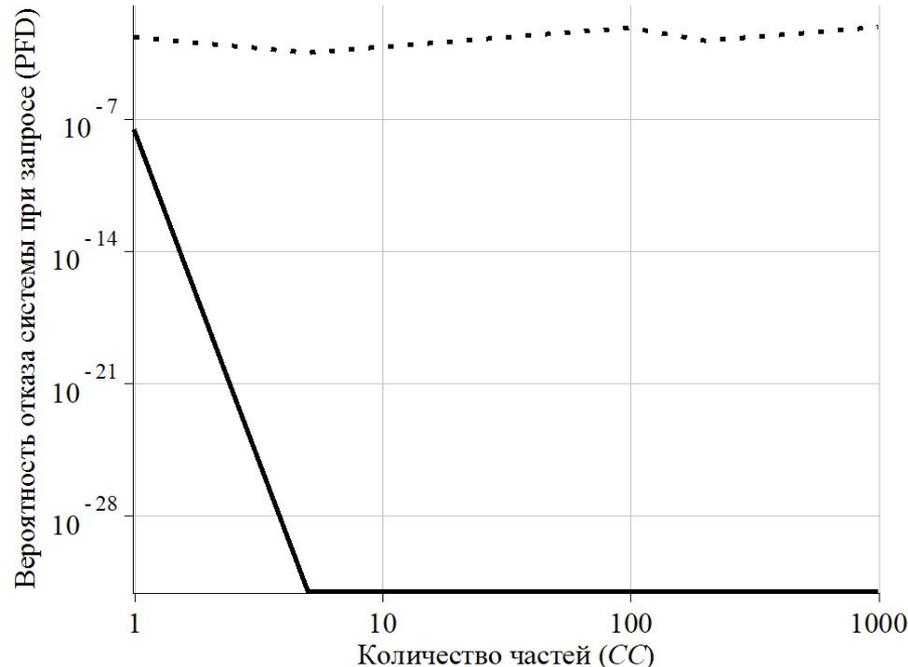
**при AFR = 0.1 в среднем в 3.36 раза;**

**при AFR = 0.01 в среднем в 2.38 раза**

# СРАВНЕНИЕ МОДЕЛЕЙ РАСПРЕДЕЛЕННОГО ХРАНЕНИЯ ДАННЫХ НА ОСНОВЕ РЕПЛИКАЦИИ И ИЗБЫТОЧНОЙ СОК



• • Стандартная модель с резервированием, AFR = 0.1  
— Предложенная модель с разделением данных, AFR = 0.1



• • Стандартная модель с резервированием, AFR = 0.01  
— Предложенная модель с разделением данных, AFR = 0.01

Условия:  $PFD \leq 10^{-2}$ , избыточность ИСОК равна избыточности при резервировании

**Преимущество ИСОК по сравнению с резервированием:**

при  $AFR = 0.1$  – в среднем на 3 порядка, с  $3.6 \cdot 10^{-3}$  до  $3 \cdot 10^{-6}$ ;

при  $AFR = 0.01$  – PFD распределенной СХД на основе ИСОК стремится к нулю

# Заключение

В настоящее время абсолютное большинство распределенных систем хранения данных (РСХД) используют репликацию для обеспечения отказоустойчивости, несмотря на высокий уровень избыточности хранения. Это связано с простотой реализации данного подхода.

Повышение отказоустойчивости и снижение эксплуатационных расходов является основной причиной, по которой пользователи РСХД заинтересованы в переходе к отказоустойчивому разделению данных. Наиболее существенные практические результаты в направлении замены репликации алгоритмами отказоустойчивого разделения данных достигнуты такими компаниями как Facebook\* и Microsoft Azure\*\*, что подчеркивает актуальность подобных исследований и их высокий научный и практический потенциал.

\* Muralidhar, S. f4: Facebook's Warm BLOB Storage System / S. Muralidhar, W. Lloyd, S. Roy [et al.] // Operating Systems Design and Implementation (OSDI): Proceedings of the 11th USENIX Symposium. – Broomfield, Colorado, USA: ACM Press, 2014. – P. 383-398.

\*\* Huang, C. Erasure coding in windows azure storage / C. Huang, H. Simitci, Y. Xu [et al.] // Annual Technical Conference: Proceedings of USENIX Conference. – Boston, Massachusetts, USA: USENIX Association, 2012. – P. 15-26.

# Спасибо за Внимание!

**Назаров Антон Сергеевич**

к.т.н., м.н.с. УНЦ «Вычислительная математика и  
параллельное программирование на супер ЭВМ»  
ФГАОУ ВО «Северо-Кавказский федеральный университет»

тел. +79187780891

e-mail: [kapitoshking@mail.ru](mailto:kapitoshking@mail.ru)