National Research
Tomsk
State
University

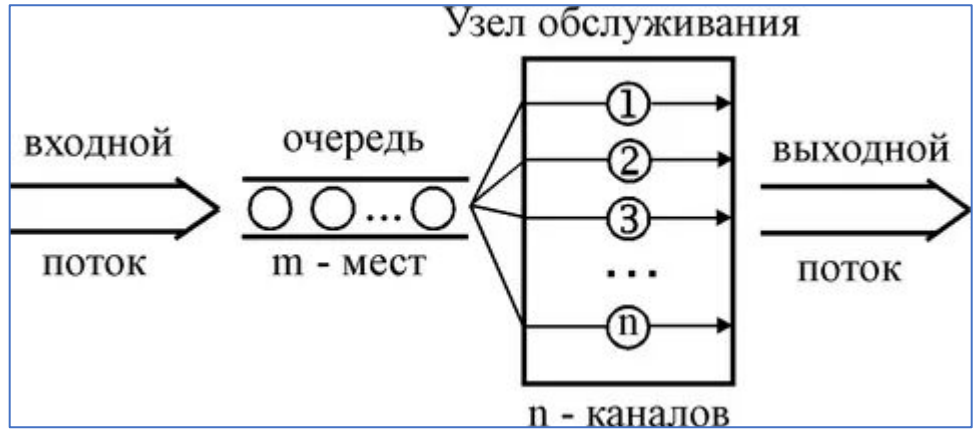
ТЕОРИЯ МАССОВОГО ОБСЛУЖИВАНИЯ В МОДЕЛИРОВАНИИ ОБЛАЧНЫХ УЗЛОВ

Иван Лапатин, Анатолий Назаров
Лаборатория теории массового обслуживания и теории телетрафика

Department of Probability Theory
and
Mathematical Statistics

План доклада

1. Немного об облачных узлах
2. Выбор моделей массового обслуживания
3. Методы исследования
4. Полезные характеристики



1. Немного об облачных узлах

1. **Немного о облачных узлах**
2. Выбор моделей массового обслуживания
3. Методы исследования
4. Полезные характеристики

Вычислительные узлы

Что можно так называть:

- Локальные компьютеры и серверы

- Облачные узлы или кластеры

Популярность облачных услуг

Что получает потребитель:

- Нет необходимости приобретать дорогостоящее оборудование
- Возможность быстро адаптировать используемые ресурсы под текущие задачи
- Задачи по обслуживанию и размещению оборудования лежат на провайдере
- Доступ к сервисам возможен из любой точки на планете, где есть интернет

Что хочет потребитель:

Платить меньше

Удобные интерфейсы

Получать хорошую скорость работы

Взгляд со стороны провайдера

Какие проблемы нужно решать

Покупка, размещение и обслуживание оборудования (**финансовые вложения**)

Необходимо избегать простоя работы «железа» (**утилизация оборудования**)

Необходимо обеспечивать определенный уровень качества предоставляемых услуг (**качество обслуживания**)

Текущее положение дел

Около 75% времени работы пиковая утилизация CPU не превышает 15%

70% энергопотребления приходится на работу в режиме ожидания

Увеличение утилизации даже на 1% приводит к значительной финансовой выгоде

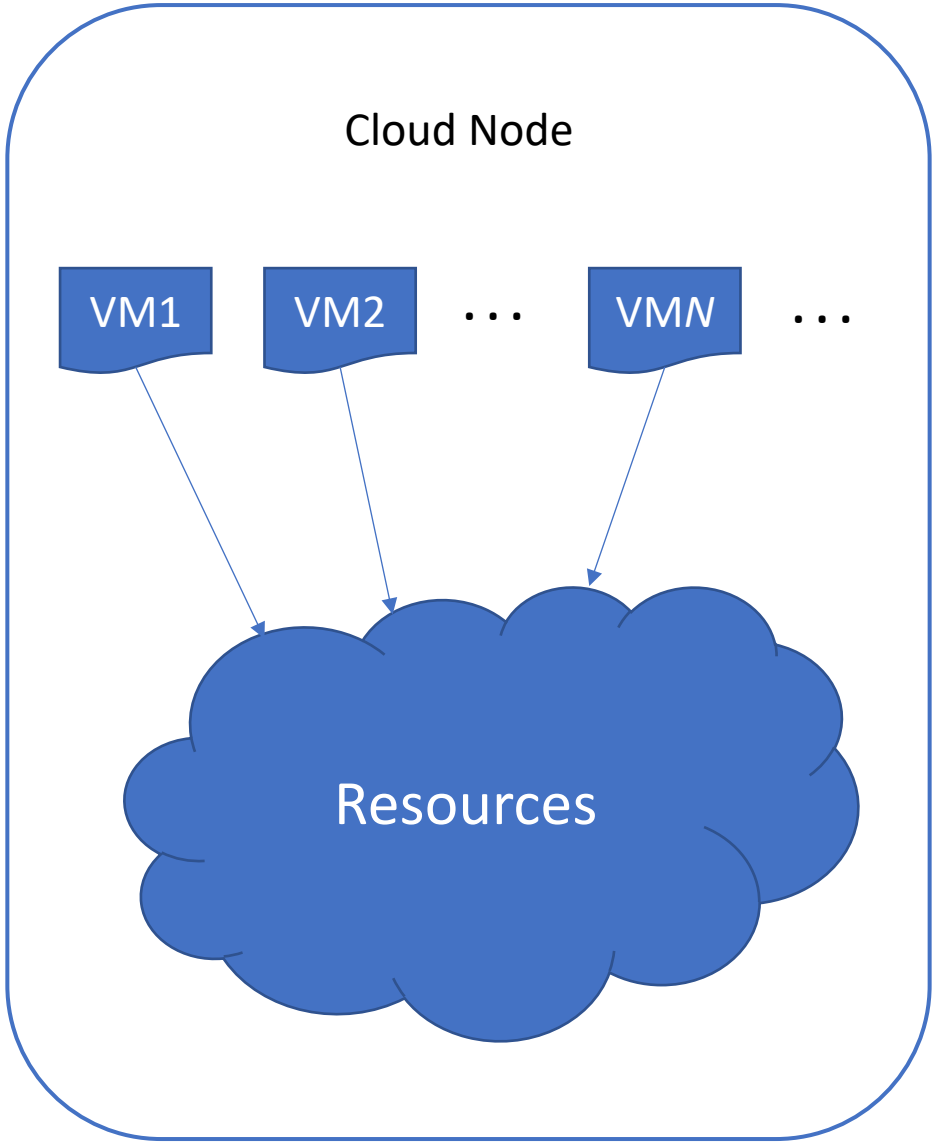
Что делает (пытается) провайдер:

- Повышать утилизацию (Utilization) таким образом, чтобы метрики (SLO – Service Level Objective) качества обслуживания (QoS – Quality of service) оставались в заданных рамках (SLI – Service Level Indicator)

Взгляд со стороны провайдера

Характерные особенности работы вычислительного узла

- Ресурсы ограничены
- Моменты возникновения запросов на использование ресурсов от пользователей имеют стохастическую природу
- Количество необходимых ресурсов для обработки запроса тоже не детерминированы
- При одновременном исполнении разных задач может возникать конкуренция за ресурсы
- Конкуренция за ресурсы приводит к снижению производительности и ухудшению качества обслуживания



Виды ресурсов

Можно выделить специфику совместного использования разных ресурсов

Процессор (CPU)

Оперативная память (DRAM)

Хранилище (Storage)

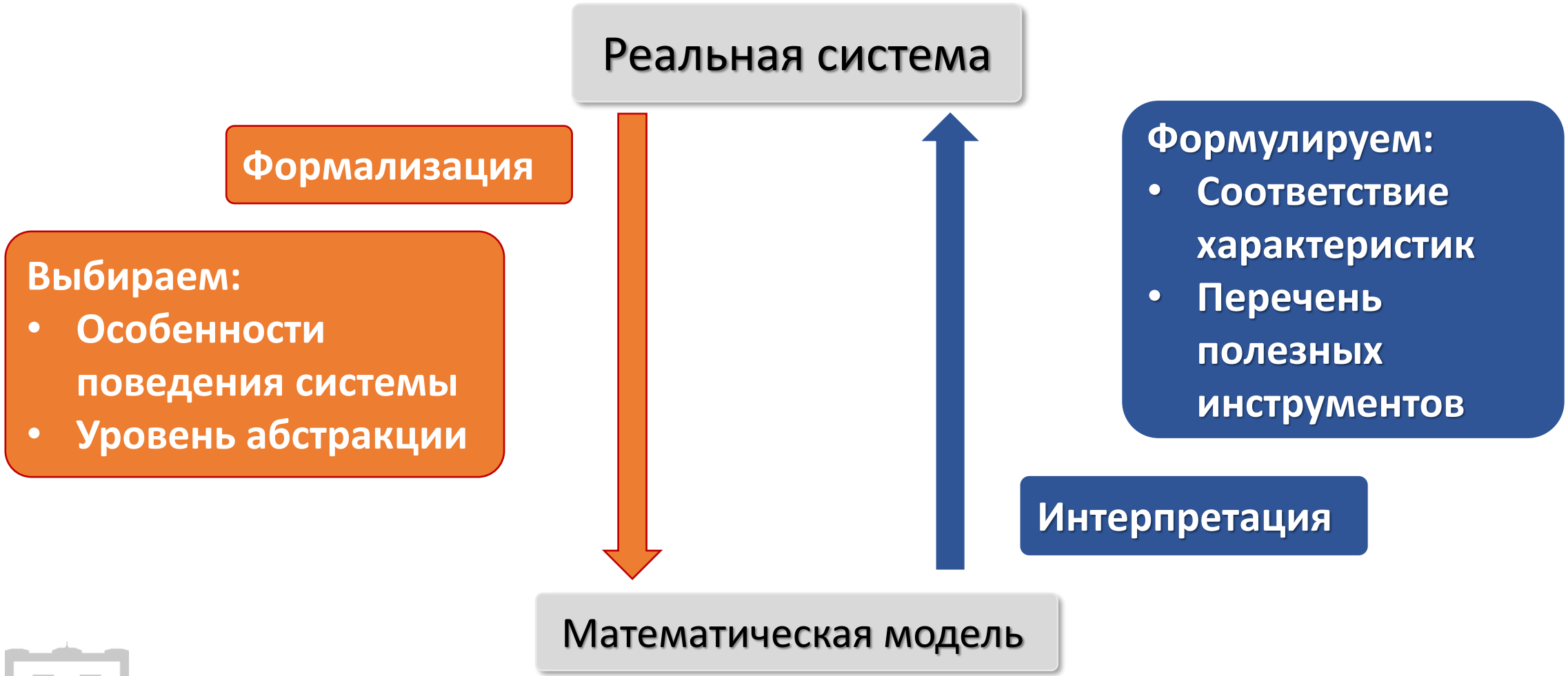
Система ввода/вывода (I/O)

Потребляемая энергия (Power)

2. Выбор моделей массового обслуживания

1. Немного об облачных узлах
2. **Выбор моделей массового обслуживания**
3. Методы исследования
4. Полезные характеристики

Математическое моделирование



Уровень абстракции - ВМ

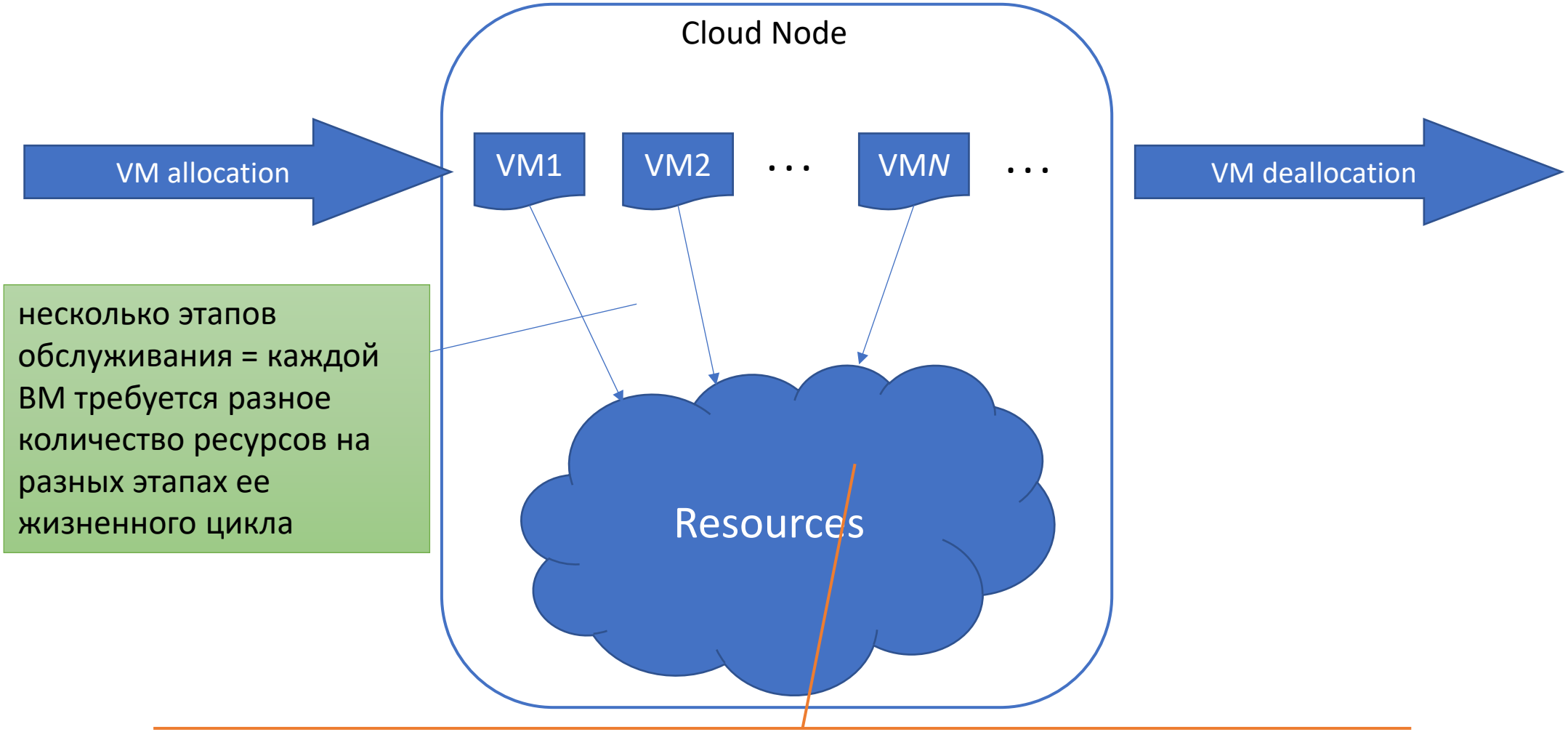
Основные особенности:

- Стохастическая природа
- Одновременное исполнения множества задач
- ВМ работает в разных режимах
- Возможное снижение производительности

Как учитываем:

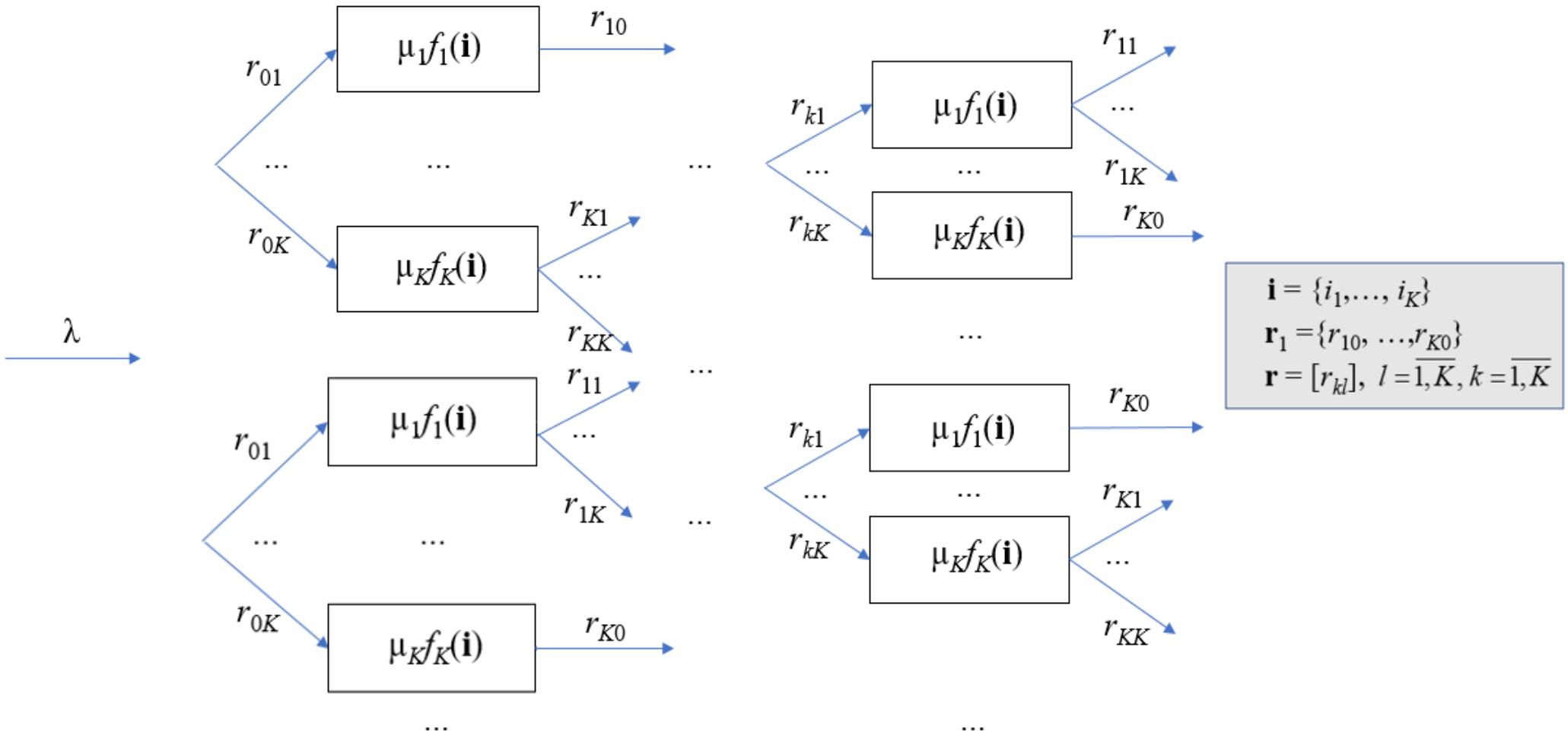
- Модели случайных потоков, времен обслуживания и матрицы маршрутизации
- Модели с неограниченным числом приборов
- Модели многофазные или модели сети
- Функция снижения скорости обслуживания от числа ВМ

Открытая модель облачного узла

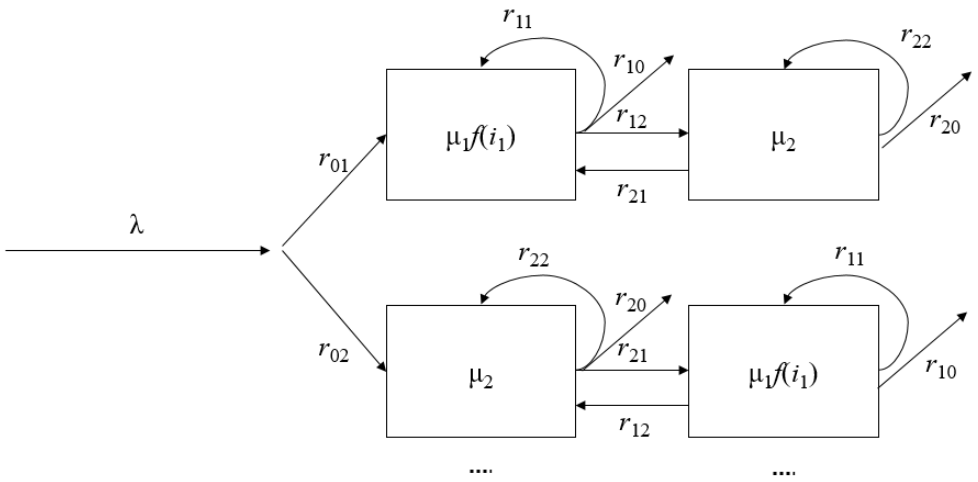


Конфликты при конкуренции за ресурсы приводят к снижению производительности=> растет время обслуживания = снижается скорость обслуживания (SRD – **Service Rate Degradation**)

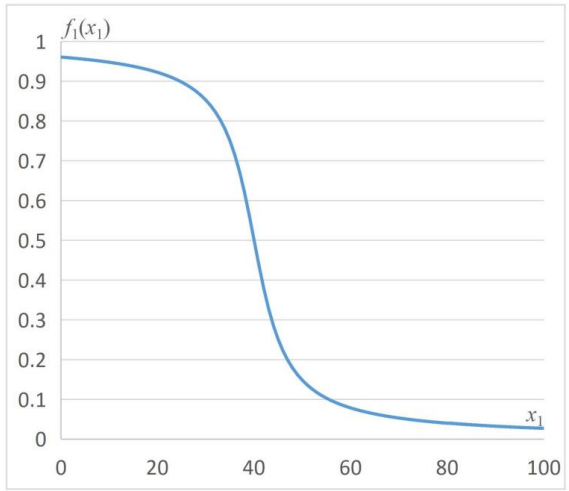
Модель узла в виде сети массового обслуживания



Базовая открытая модель узла в виде CeMO



r_{kl} – probability of transition from k-th phase to l-th, $l, k = 1, 2$
$\mu_1 f(i_1)$ – service rate on 1-th phase

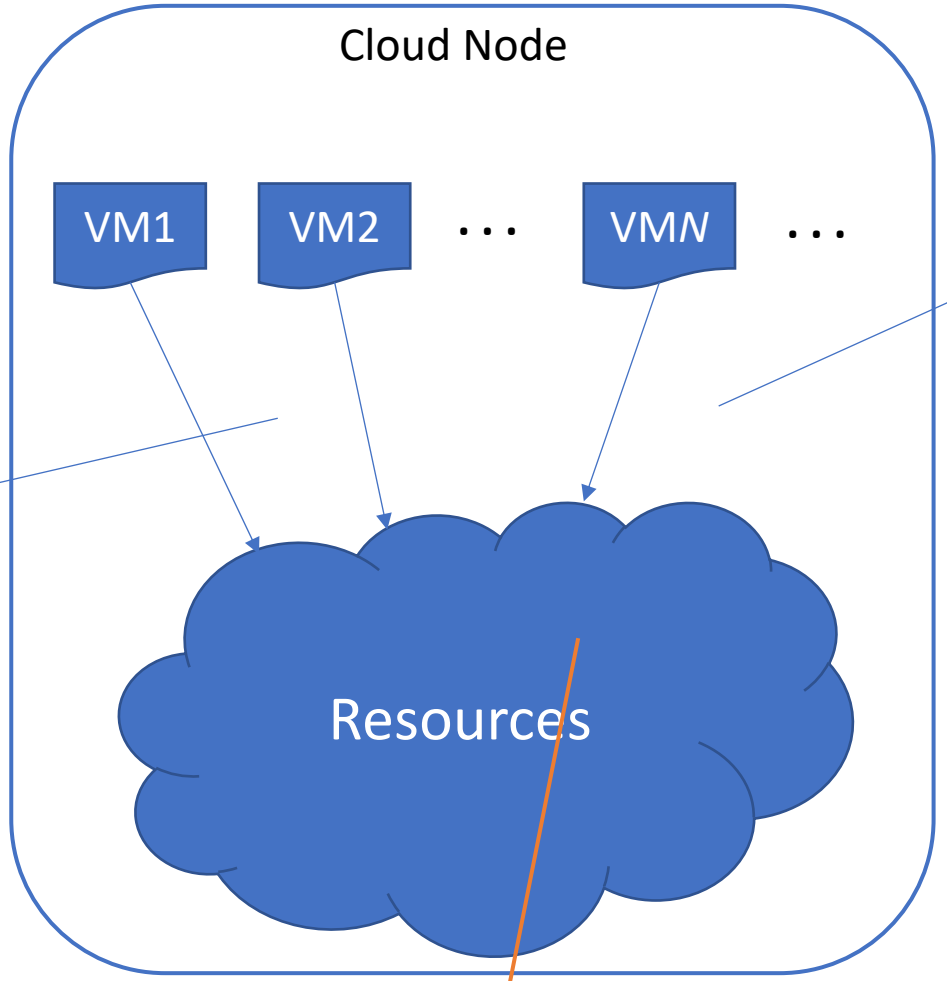


Обозначения	Значения
λ	Интенсивность запуска ВМ на узле
$\mu_1 f(i_1)$	Интенсивность обслуживания ВМ на первой фазе
μ_2	Интенсивность обслуживания ВМ на первой фазе
R	Матрица вероятностей r_{kl} переключения ВМ с k-й фазы на l-ю фазу, $k = 1, 2$, $l = 1, 2$
r_0	Вектор вероятностей r_{0k}, $k = 1, 2$
r_1	Вектор вероятностей r_{k0} того, что ВМ покидает узел после k-й фазы, $k = 1, 2$

Допущения:

- Две фазы = два режима работы ВМ (активная работа/режим ожидания)
- Количество *N* ВМ фиксировано

Закрытая модель вычислительного узла

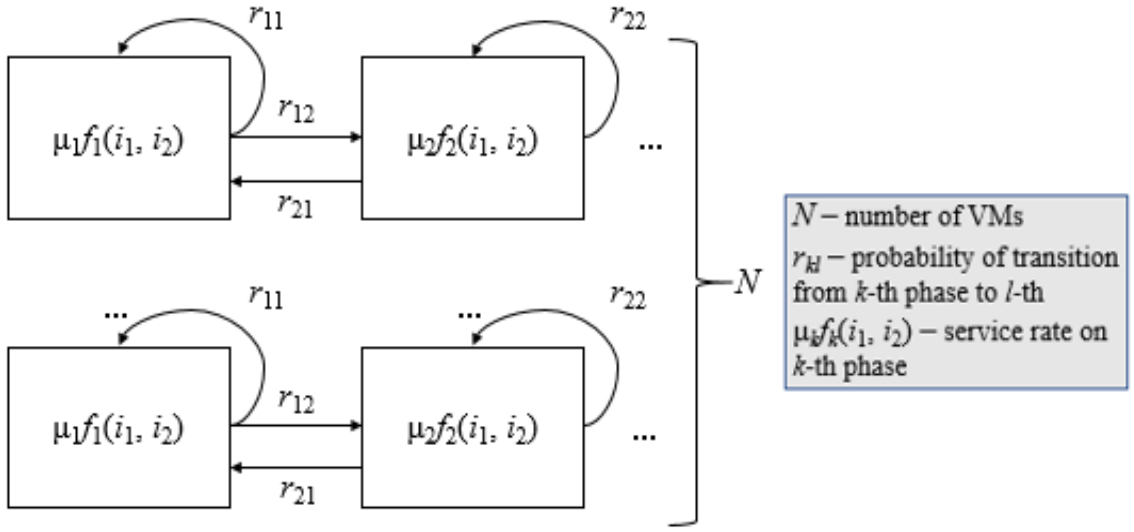


Несколько этапов обслуживания = каждой ВМ требуется разное количество ресурсов на разных этапах ее жизненного цикла

На узле запущено определенное количество ВМ, которое не меняется на протяжении длительного времени, происходит только смена режимов работы

Конфликты при конкуренции за ресурсы приводят к снижению производительности=> растет время обслуживания = снижается скорость обслуживания (SRD – **Service Rate Degradation**)

Закрытая двухфазная модель узла в виде CeMO



Обозначения	Значения
N	Число ВМ в узле
R	Матрица вероятностей r_{kl} переключения ВМ с k-й фазы на l-ю фазу, k = 1, 2, l = 1, 2
$\mu_n f_n(i_1, i_2)$	Интенсивность обслуживания ВМ на n-й фазе, n = 1, 2
$f_n(i_1, i_2)$	Функция деградации скоторости обслуживания

Объект исследования:
Совместное распределение вероятностей числа ВМ на каждой фазе
Маргинальные распределения или их числовые характеристики

Допущения:

- Две фазы = два режима работы ВМ (активная работа/режим ожидания)
- Количество *N* ВМ фиксировано

Уровень абстракции – запрос

Запрос – событие, требующее выполнить некоторый объем работы на узле, используя определенные его ресурсы

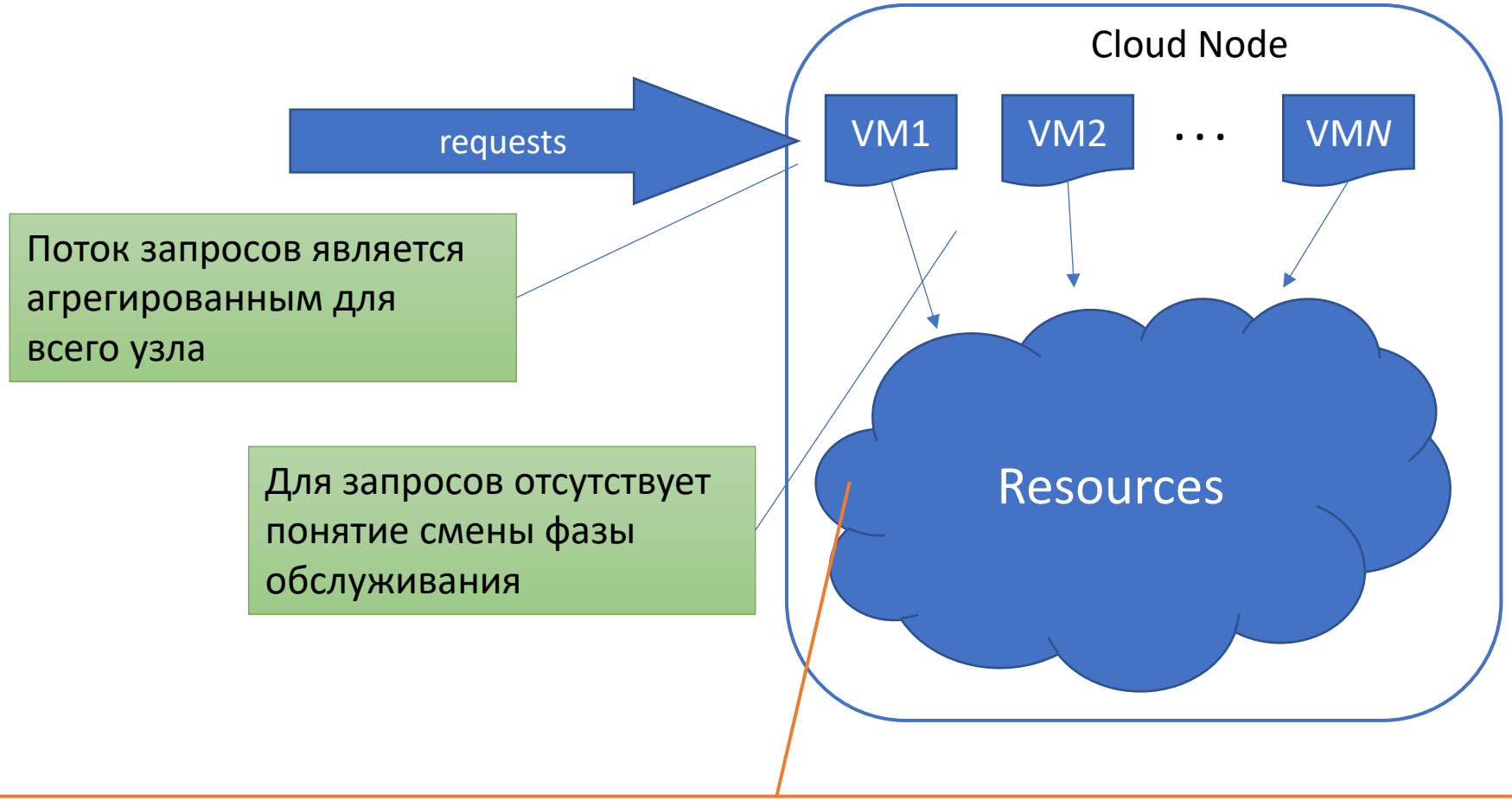
Основные особенности:

- Стохастическая природа
- Одновременное исполнение множества задач
- Снижение скорости обработки запросов при увеличении числа выполняемых запросов

Как учитываем:

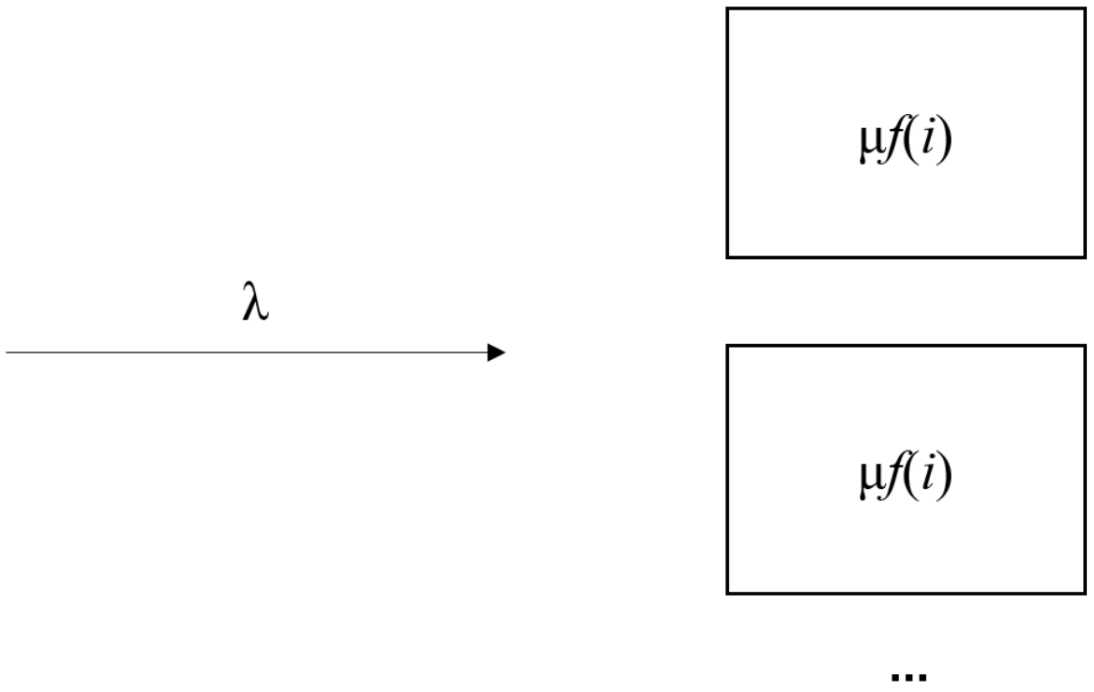
- Модели случайных потоков, времен обслуживания
- Модели с неограниченным числом приборов
- Функция деградации скорости обслуживания от числа запросов

Общий вид модели узла на уровне запросов



Конфликты при конкуренции запросов за ресурсы приводят к снижению производительности=> растет время обслуживания = снижается скорость обслуживания (SRD – **Service Rate Degradation**)

Модель узла в виде СМО с неограниченным числом приборов



Объект исследования:
Распределение вероятностей или числовые характеристики числа запросов в системе

Особенности модели:

- Одна фаза обслуживания, после обслуживания запрос покидает систему
- Количество одновременно выполняемых запросов неограниченно

Публикации по предложенным моделям

- Лапатин И.Л., Назаров А.А., Пауль С.В. Модель работы процессора в условиях конкуренции за вычислительный ресурс // Распределенные компьютерные и телекоммуникационные сети: управление, вычисление, связь (DCCN-2023): материалы XXVI Международной научной конференции (25–29 сентября 2023 г., Москва, Россия). М.: ИПУ РАН, 2023. С. 195–200. CD-R.
- Asymptotic Analysis of Two-Phase Queueing System with Service Rate Degradation and Heterogeneous Customers / E.A. Fedorova, I.L. Lapatin, Lizyura O. D., Moiseev A.N. [et al] // 2023 5th International Conference on Problems of Cybernetics and Informatics (PCI) 2023, 28-30 august 2023. [S. l.], 2023. P. 1–5. URL: https://ieeexplore.ieee.org/document/10326020 (date of access: 29.11.2023)
- Mathematical Modeling of Virtual Machine Life Cycle Using Branching Renewal Process / E.A. Fedorova, I.L. Lapatin, O.D. Lizyura, A.N. Moiseev [et al] // Communications in Computer and Information Science. 2023. Vol. 1803. P. 29–39. DOI: 10.1007/978-3-031-32990-6_3
- Queueing System with Two Phases of Service and Service Rate Degradation / E.A. Fedorova, I.L. Lapatin, O.D. Lizyura, A.N. Moiseev [et al] // Axioms. 2023. Vol. 12, № 2. Art. num. 104. URL: https://www.mdpi.com/2075-1680/12/2/104.

3. Методы исследования

1. Немного об облачных узлах
2. Выбор моделей массового обслуживания
3. **Методы исследования**
4. Полезные характеристики

Возможные методы исследования предложенных моделей

Аналитические методы

Численные методы

Имитационное моделирования

Методы для аналитического анализа

Широко применяемые методы

- Метод дополнительной переменной
- Метод расширения фазового пространства
- Метод производящей функции
- Метод характеристической функции
- И другие…

Авторские методы команды ТГУ

- Метод динамического просеивания
- Метод многомерного динамического просеивания
- Методы асимптотического анализа
- Метод асимптотически диффузионного анализа

4. Полезные характеристики

1. Немного об облачных узлах
2. Выбор моделей массового обслуживания
3. Методы исследования
4. **Полезные характеристики**

Результаты анализа математической модели

Получаем стационарное распределение вероятностей $P(i)$ числа занятых приборов в системе или на отдельной фазе

Вероятности $P(i)$ имеют смысл доли времени, когда в системе или на отдельной фазе занято i приборов

Зная распределение вероятностей, можно получить числовые характеристики – среднее, дисперсию, квантили

Что это означает для облачного узла

Стационарное распределение вероятностей $P(i)$ определяет какую долю времени на узле одновременно обрабатывалось i запросов

Если на узле имеется K ядер, то мы легко можем рассчитать утилизацию (utilization) вычислительного ресурса

Количество одновременно обрабатываемых запросов определяет уровень снижения скорости обслуживания (SLO)

Базовая открытая модель узла в виде СеМО

Объект исследования:

- Стационарное распределение вероятностей $i_1(t)$ и $i_2(t)$ числа ВМ на первой и второй фазах $P(i_1, i_2) = \mathrm{P}\{i_1(t) = i_1, i_2(t) = i_2\}$

$$p(x_1, x_2) \approx \frac{1}{2\pi\sqrt{K_{11}K_{22}(1-\eta^2)}} \exp\left\{-\frac{1}{2(1-\eta^2)}\left[\frac{(x_1-\kappa_1)^2}{K_{11}} - \eta\frac{2(x_1-\kappa_1)(x_2-\kappa_2)}{\sqrt{K_{11}K_{22}}} + \frac{(x_2-\kappa_2)^2}{K_{22}}\right]\right\} \quad (1)$$

$$\kappa_2 = \frac{r_{12}\mu_1\kappa_1 f_1(\kappa_1) + \lambda r_{02}}{\mu_2} \qquad \kappa_1 f_1(\kappa_1) = \frac{\lambda(r_{01} + r_{02}r_{21})}{\mu_1(1-r_{21}r_{12})} \quad (2)$$

$$K_{11} = \frac{b_1 - K_{12}a_{12}}{a_{11}} \quad K_{22} = \frac{b_2 - K_{12}a_{21}}{a_{22}} \quad K_{12} = \frac{b_{12}a_{11}a_{22} - a_{21}b_1a_{22} - a_{12}b_2a_{11}}{a_{11}^2a_{22} - a_{21}a_{12}a_{22} - a_{12}a_{21}a_{11} + a_{22}^2a_{11}} \quad \eta = \sqrt{\frac{K_{12}}{K_{11}K_{22}}} \quad (3)$$

$$a_{11} = \mu_1 \frac{\partial}{\partial x_1}\{x_1 f_1(x_1)\}\Big|_{x_1=\kappa_1} \quad a_{12} = -r_{21}\mu_2 \quad a_{21} = -r_{12}\mu_1 \frac{\partial}{\partial x_1}\{x_1 f_1(x_1)\}\Big|_{x_1=\kappa_1} \quad a_{22} = \mu_2 \quad (4)$$

$$b_1 = \mu_1\kappa_1 f_1(\kappa_1) \quad b_2 = \mu_2\kappa_2 \quad b_{12} = r_{12}\mu_1\kappa_1 f_1(\kappa_1) + r_{21}\mu_2\kappa_2$$

Базовая открытая модель узла в виде СеМО

Объект исследования :

- Маргинальные распределения $P(i_1)$ и $P(i_2)$ числа $i_1(t)$ и $i_2(t)$ ВМ на каждой фазе

$$p(x_1) = \frac{1}{\sqrt{2\pi K_{11}}} \times \exp\left\{-\frac{(x_1-\kappa_1)^2}{2K_{11}}\right\} \qquad p(x_2) = \frac{1}{\sqrt{2\pi K_{22}}} \times \exp\left\{-\frac{(x_2-\kappa_2)^2}{2K_{22}}\right\} \tag{5}$$

$$\kappa_2 = \frac{r_{12}\mu_1\kappa_1 f_1(\kappa_1) + \lambda r_{02}}{\mu_2} \qquad \kappa_1 f_1(\kappa_1) = \frac{\lambda(r_{01} + r_{02}r_{21})}{\mu_1(1-r_{21}r_{12})} \tag{6}$$

$$K_{11} = \frac{b_1 - K_{12}a_{12}}{a_{11}} \quad K_{22} = \frac{b_2 - K_{12}a_{21}}{a_{22}} \tag{7}$$

$$a_{11} = \mu_1 \frac{\partial}{\partial x_1}\{x_1 f_1(x_1)\}\Big|_{x_1=\kappa_1} \quad a_{12} = -r_{21}\mu_2 \quad a_{21} = -r_{12}\mu_1 \frac{\partial}{\partial x_1}\{x_1 f_1(x_1)\}\Big|_{x_1=\kappa_1} \quad a_{22} = \mu_2$$

$$b_1 = \mu_1\kappa_1 f_1(\kappa_1) \quad b_2 = \mu_2\kappa_2 \tag{8}$$

Спасибо за внимание!

Tomsk State University

36, Lenin Avenue, Tomsk, 634050, Russia
www.tsu.ru

Department of Probability Theory and Mathematical Statistics
csi.tsu.ru
vk.com/tv_ms