

УДК 658.012:681.324

А.В. Скатков, профессор, д-р техн. наук,

И.А. Балакирева, доцент, канд. техн. наук

ул. Университетская, 33, г. Севастополь, Украина, 99053

E-mail: kvt.sevntu@gmail.com

УПРАВЛЕНИЕ ВЫЧИСЛИТЕЛЬНОЙ МОЩНОСТЬЮ УЗЛА ПРИ КРИТИЧЕСКИХ СОБЫТИЯХ

Рассмотрена задача управления мощностью вычислительного узла в условиях критической нагрузки. Анализируется стратегия принятия решений о введении в работу дополнительного вычислительного канала в случае превышения нагрузкой узла критического значения. Эффективность предлагаемого подхода исследована в ходе вычислительного эксперимента.

Ключевые слова: ИТ-сервис, вычислительный узел, критическая рабочая нагрузка, стратегия управления, инцидент.

Развитие информационных технологий (ИТ) предъявляет ряд требований к организации управления ИТ-инфраструктурой. Управленческие решения по поддержке и обслуживанию такой инфраструктуры обеспечиваются ИТ-сервисами.

Эффективность функционирования ИТ-сервиса, согласно сведениям ITIL [1], прежде всего, зависит от того, как в рамках данной ИТ-инфраструктуры решаются вопросы, связанные с управлением уровнем ИТ-услуг, их непрерывностью и доступностью, финансами, а также мощностями, которыми располагает ИТ-инфраструктура.

Управление мощностями представляет собой процесс, связанный с оптимизацией расходов, времен приобретения и размещения ИТ-ресурсов. Здесь же решаются проблемы управления производительностью инфраструктуры и планированием ее мощностей. Согласно Соглашению об Уровне Обслуживания (SLA), определение необходимого объема технических средств и управление нагрузкой являются важными вопросами при решении задачи обеспечения непрерывности ИТ-услуг.

В связи с этим особый интерес представляет собой процесс управления мощностями ИТ-инфраструктуры при пиковых нагрузках и критических событиях – инцидентах. Если при прогнозируемых пиковых нагрузках в течение того или иного периода времени (сутки, неделя, год) заранее планируется увеличение мощностей ИТ-инфраструктуры, то критические пиковые нагрузки предполагают немедленное изменение конфигурации инфраструктуры для обеспечения непрерывности ИТ-сервиса.

Цель статьи заключается в определении стратегии принятия решений при управлении мощностью вычислительного узла (ВУ) за счет введения дополнительного вычислительного канала в условиях критического изменения рабочей нагрузки.

Задача определения стратегии управления ВУ в условиях критической нагрузки рассматривается на примере ВУ с буферной памятью известного объема на входе узла. ВУ является компонентом коммуникационной сети.

В ВУ поступает поток заданий, переменный во времени, с интенсивностью $\lambda(t)$. Производительность процессора вычислительного узла постоянна и составляет μ заданий в единицу времени. Объем оперативной памяти узла, предназначенный для хранения заданий, ожидающих обслуживания, ограничен и является достаточным, если интенсивность потока заданий не превышает некоторое критическое значение: $\lambda(t) \leq \lambda_{кр}$.

Если возникает ситуация, такая, что интенсивность потока заданий превышает некоторое критическое значение, то ее можно характеризовать как инцидент или критическое событие. Подобные события в ВУ приводят к потере данных и ставят под вопрос выполнение требований непрерывности ИТ-сервиса.

В такой ситуации целесообразно определить стратегию включения в конфигурацию ВУ дополнительных вычислительных каналов. Это позволяет увеличить производительность данного узла и, тем самым, предотвратить инцидент, связанный с потерей данных.

Для определения правила управления мощностью ВУ необходимо установить его конфигурацию, оптимальную с точки зрения ожидаемых затрат на ИТ-ресурсы. Очевидно, что, чем больше производительность процессора, тем меньшие затраты необходимы для организации хранения заданий, ожидающих обслуживания. С другой стороны, увеличение производительности связано с повышением затрат на оборудование за счет включения в конфигурацию узла дополнительных вычислительных каналов. Таким образом, эффективность работы ВУ можно повысить за счет минимизации очереди заданий, поступающих на вход ВУ и использования дополнительных вычислительных каналов, а также за счет увеличения коэффициента использования ресурсов ВУ, что представимо несколькими скалярными критериями соответственно $I_{оч} \rightarrow \min$, $\mu \rightarrow \min$, и $\rho \rightarrow \max$.

В такой ситуации целесообразно для определения стратегии управления построить агрегированный критерий, который выразит предпочтения лица, принимающего решения (ЛПР), и позволит определить значения параметров вычислительного узла, минимизирующие затраты на используемые ИТ-ресурсы, например

$$E(l_{оч}, \mu) = c_l l_{оч}(\mu, \lambda) + c_p(1 - \rho) + c_\mu \mu \rightarrow \min, \quad (1)$$

где c_l , c_p и c_μ – весовые коэффициенты, выражающие соответственно стоимости эксплуатации одного места в очереди ожидания заданий на обслуживание, штрафа за недогрузку вычислительного канала и за используемые мощности ВУ, причем $c_l + c_p + c_\mu = 1$; $l_{оч} = l_t / \bar{l}$ – относительная длина очереди заданий на обслуживание; ρ – коэффициент использования ВУ; $\mu = \mu_t / \bar{\mu}$ – относительная производительность вычислительного канала; $\lambda = \lambda_t / \bar{\lambda}$ – относительная интенсивность поступления заданий на обработку в ВУ.

В приведенных соотношениях в числителях представлены текущие значения, а в знаменателях – средние значения соответствующих параметров. Далее по тексту и на рисунках речь идет об относительных значениях производительности вычислительных каналов, интенсивности входного потока заданий и длины очереди заданий, ожидающих выполнения в вычислительном узле.

На основании решения задачи оптимизации в соответствии с критерием (1) можно определить оптимальную производительность ВУ, обеспечивающую требование минимизации затрат при эксплуатации узла при непрерывном изменении его производительности.

Как правило, изменение производительности ВУ осуществляется скачкообразно путем подключения дополнительного вычислительного канала, т.е.

$$\mu = \begin{cases} \mu_1, & \text{если } l_{оч} \leq l_{кр} \\ \mu_2, & \text{если } l_{оч} > l_{кр} \end{cases}, \quad (2)$$

где μ_2 – суммарная производительность основного и дополнительного вычислительных каналов, $l_{кр}$ – максимально возможная длина очереди заданий на входе ВУ при производительности равной μ_1 .

Таким образом, в общем случае производительность ВУ является переменным параметром.

В условиях переменной производительности и на основании (2) критерий эффективности функционирования ВУ можно представить следующим образом

$$E(u, (l_{кр}, \lambda, \mu)) \rightarrow \min, \quad (3)$$

где u – переменная управления, в зависимости от значения которой осуществляется изменение производительности ВУ:

$$u = \begin{cases} 0, & \text{если } l_{оч} \leq l_{кр}; \\ 1, & \text{если } l_{оч} > l_{кр}. \end{cases} \quad (4)$$

На основании решения задачи оптимизации в соответствии с критерием (3), можно определить требуемую производительность и длину очереди заданий на входе ВУ, при которых будет обеспечиваться требование минимизации затрат при эксплуатации узла в условиях критической нагрузки.

Используя правило принятия решений, необходимо изменять производительность ВУ в зависимости от длины очереди заданий, ожидающих выполнения. При наступлении критического события, при котором длина очереди превысит уровень $l_{кр}$, необходимо повысить мощность ВУ, и, тем самым, предотвратить инцидент, связанный с переполнением буферной памяти на входе и, как следствие, с потерей данных. При нормализации длины очереди необходимость в использовании дополнительных мощностей ВУ теряет смысл и дополнительный канал можно отключить.

Исследование системы управления мощностью ВУ осуществлялось на основании модели, построенной в системе Simulink [2]. Основой моделирования являются системы дифференциальных уравнений (ДУ) Колмогорова [3]. Рассматривается модель, имеющая три возможных состояния S_0 , S_1 и S_2 .

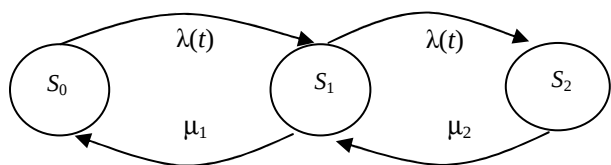


Рисунок 1 – Граф состояний ВУ при нестационарной нагрузке

На вход вычислительного узла подается поток заявок с переменной во времени интенсивностью $\lambda(t)$. Размеченный граф состояний ВУ изображен на рисунке 1. Состояния имеют следующий смысл: S_0 – ВУ свободен, S_1 и S_2 – соответственно означают, что в ВУ заняты либо один вычислительный канал, либо два канала одновременно с производительностями, указанными на графе.

Размеченному графу состояний ВУ соответствует система ДУ:

$$\begin{cases} \frac{dP_0(t)}{dt} = \mu_1 P_1(t) - \lambda(t) P_0(t); \\ \frac{dP_1(t)}{dt} = \lambda(t) P_0(t) + \mu_2 P_2(t) - (\lambda(t) + \mu_1) P_1(t); \\ \frac{dP_2(t)}{dt} = \lambda(t) P_1(t) - \mu_2 P_2(t). \end{cases} \quad (5)$$

Для функций, входящих в (5), в каждый момент времени выполняется условие $P_0(t) + P_1(t) + P_2(t) = 1$. Начальные условия для системы ДУ отражают состояния ВУ в начальный момент и для системы (5) считаются известными $P_0(t) = 1$, $P_1(t) = P_2(t) = 0$.

На рисунке 2 представлена укрупненная схема модели функционирования ВУ с переменной мощностью.

Модель содержит блок формирования входных ресурсов [2] – поток заданий с переменной интенсивностью $\lambda(t)$. Блоки μ_1 и μ_2 имитируют функционирование вычислительного узла при соответствующей производительности. Управление мощностью ВУ осуществляется переключателем switch, в зависимости от длины очереди заданий, ожидающих выполнения, – $l_{кр}$.

На основании предлагаемой модели можно определить такие параметры ВУ, как коэффициент загрузки, вероятности использования вычислительных каналов, среднюю длину очереди и другие, как функции производительности ВУ и интенсивности входного потока, которые могут быть интересны при решении той или иной оптимизационной задачи, например, задачи (1).

Для исследования сформулированной задачи (3), (4) управления мощностью ВУ были проведены вычислительные эксперименты. Эксперимент преследовал следующие цели:

- определить критическое значение длины очереди заданий в соответствии с критерием (1) и выяснить влияние параметров на значение критерия;
- изучить характер изменения средней длины очереди заданий на обслуживание в условиях переменной производительности ВУ.

Эксперимент проводился при условии, что изменение интенсивности входного потока осуществляется линейно в соответствии с выражением $\lambda(t) = \lambda_{0i} + a_i t$, где i – номер соответствующего отрезка времени за рассматриваемый период. Характер изменения во времени интенсивности входного потока изображен на рисунке 3.

Шкала времени на этом рисунке и всех последующих представлена в условных единицах

$$t = \frac{t_{\text{мод}}}{t_{\text{макс}}} \cdot 100\%, \text{ где } t_{\text{мод}} \text{ и } t_{\text{макс}}^{\text{max}} \text{ — соответственно}$$

текущее и максимальное реальное моделируемое время процесса изменения мощности ВУ.

В качестве примера рассматривался ВУ с производительностью $\mu_1 = \mu$. В случае инцидента, связанного с превышением длины очереди уровня $l_{кр}$, производительность узла составляла $\mu_2 = 2\mu$.

1. *Определение стратегии управления производительностью ВУ.* На основании разработанной модели и заданных параметров вычислены значения критерия эффективности (1). На рисунке 4, а представлен характер изменения функции, описывающей затраты на эксплуатацию ВУ, а также его составляющие, связанные с затратами на организацию очереди заданий, ожидающих обслуживания и с ресурсными затратами ВУ.

В ходе аналитического исследования стратегии управления мощностью ВУ выяснено, что при фиксированном значении интенсивности входного потока заданий $\lambda(t) = \lambda$ критерий (1) достигает минимального значения при определенной конфигурации ВУ производительности μ^* .

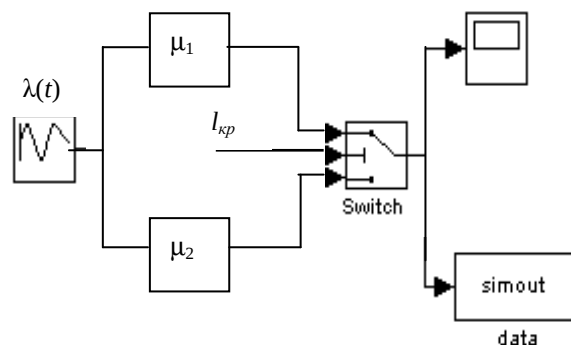


Рисунок 2 – Модель управления мощностью ВУ

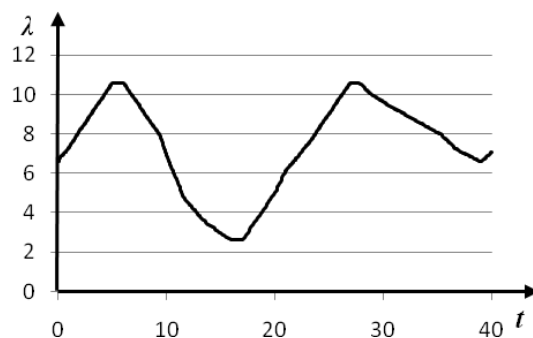


Рисунок 3 – Тестовая интенсивность входного потока заданий

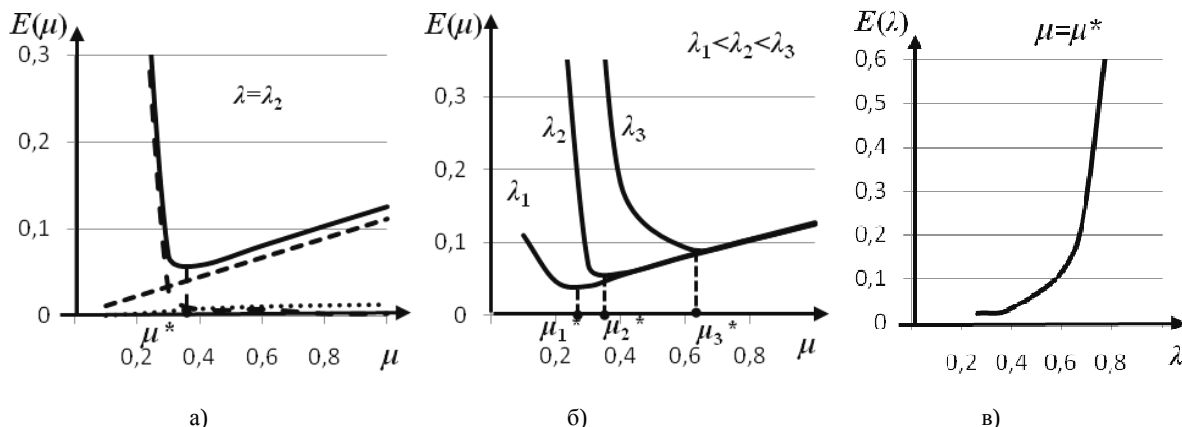


Рисунок 4 – Критерий эффективности управления мощностью ВУ: а) составляющие затрат при использовании ИТ-ресурсов ВУ в случае переменной производительности; б) уровень затрат на ИТ-ресурсы при изменении интенсивности потока заданий на входе ВУ при переменной производительности; в) уровень затрат на ИТ-ресурсы при переменной интенсивности входного потока и оптимальной производительности

Координата μ^* , соответствующая оптимальному значению производительности ВУ, определяется параметрами решаемой задачи. На рисунке 4,б изображены графики критерия (1) в зависимости от интенсивности потока заданий на входе ВУ. Значения весовых коэффициентов критерия c_b , c_p и c_μ выражают отношение ЛПР к использованию ИТ-ресурсов ВУ и оказывают влияние на характер изменения критерия, а также на оптимальное значение производительности.

Итак, установлено, что при постоянной интенсивности входного потока заявок λ существует оптимальное значение μ^* производительности ВУ. Этим параметрам соответствует единственное значение длины очереди l_{kp} [3]. При пиковых значениях интенсивности входного потока $\lambda(t) > \lambda$ длина очереди, а, следовательно, и значение критерия эффективности (1) резко возрастает, что отображено на рисунке 4,в. В такой ситуации для обеспечения непрерывности ИТ-сервиса необходимо либо увеличивать объем памяти для хранения заданий в очереди, либо повысить производительность ВУ, например, за счет подключения дополнительного вычислительного канала.

В ходе вычислительного эксперимента исследовался характер изменения критерия эффективности (1) при увеличении интенсивности входного потока заданий. Производительность ВУ при этом изменялась в соответствии с (2) в зависимости от установленной критической длины очереди. Графическая интерпретация эксперимента в виде уровня затрат на ИТ-ресурсы показана на рисунке 5.

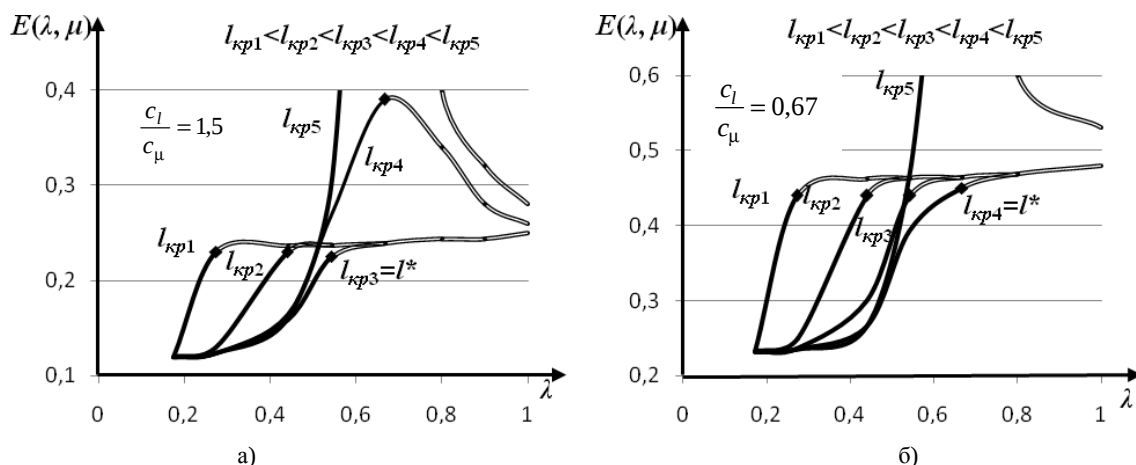


Рисунок 5 – Критерий эффективности управления мощностью ВУ с переменной производительностью при различных размерах штрафов за использование ИТ-ресурсов: а) максимальный штраф за хранение заданий на входе ВУ; б) максимальный штраф за используемые мощности

В условиях изменения интенсивности входного потока заданий на входе ВУ и производительности узла μ_1 средняя длина очереди заявок на входе возрастает. Если длину очереди, при которой осуществляется подключение дополнительного вычислительного канала l_{kp} , принять равной $l_{kp} = 0$, то это будет означать, что ВУ использует как основной, так и дополнительный вычислительные каналы. Потери в этом случае в основном связаны с производительностью узла. Если $l_{kp} = \infty$, то

дополнительный канал не используется, и потери определяются в основном за счет увеличения длины очереди заданий на входе ВУ (2). Если критическая длина очереди принимает конечное значение l_{kpi} (эксперимент проводился при $i=1,5$), то при достижении длиной очереди указанного значения осуществляется подключение дополнительного вычислительного канала. При этом производительность вычислительного узла становится равной μ_2 . На рисунке 5 затраты, связанные с использованием основного вычислительного канала производительностью μ_1 , изображены одинарной линией. Тот же показатель, но при использовании, как основного, так и дополнительного каналов с суммарной производительностью μ_2 , изображен двойной линией.

Для того, чтобы воспользоваться критерием эффективности (3) разобьем рассматриваемый интервал изменения интенсивности входного потока заданий $\lambda_{\max} - \lambda_{\min}$ на отрезки $\Delta\lambda$ и вычислим значение накопленных (интегральных), затрат на использование ИТ-ресурсов в наблюдаемом интервале изменения интенсивности входного потока

$$E_n(l_{kp}, \lambda, \mu) = \sum_{j=1}^n E(l_{kp}, \lambda_j, \mu_1) \Delta\lambda, \quad (6)$$

где λ_t — текущее значение интенсивности входного потока заданий; n — число отрезков разбиения интервала $\lambda_{\max} - \lambda_{\min}$.

При $\Delta\lambda \rightarrow 0$ выражение (6) примет вид:

$$E_n(l_{kp}, \lambda, \mu) = \int_{\lambda_{\min}}^{\lambda_{\max}} E(l_{kp}, \lambda_t, \mu_1) d\lambda. \quad (7)$$

Размер интегральных затрат на использование ИТ-ресурсов в соответствии с (7) изображен на рисунке 6, а. Закрашенные области соответствуют затратам при различных значениях критической длины очереди, при этом штраф за используемые мощности превышает штраф за хранение заданий в очереди (рисунк 5, б). На рисунке видно, что при значении критической длины очереди заданий $l_{kp4} > l_{kp1}$ интегральные затраты меньше. Поэтому при заданных параметрах ВУ предпочтительно подключать дополнительный вычислительный канал, когда длина очереди заданий на входе ВУ достигнет уровня l_{kp4} .

Характер изменения интегрального показателя потерь в зависимости от критической длины очереди и от предпочтений ЛПР на используемые ИТ-ресурсы ВУ изображен на рисунке 6, б. Выяснено, что существует такое значение критической длины очереди, при котором необходимо увеличить производительность ВУ с целью минимизации затрат на используемые ресурсы узла. Оптимальное значение критической длины очереди определяется параметрами задачи.

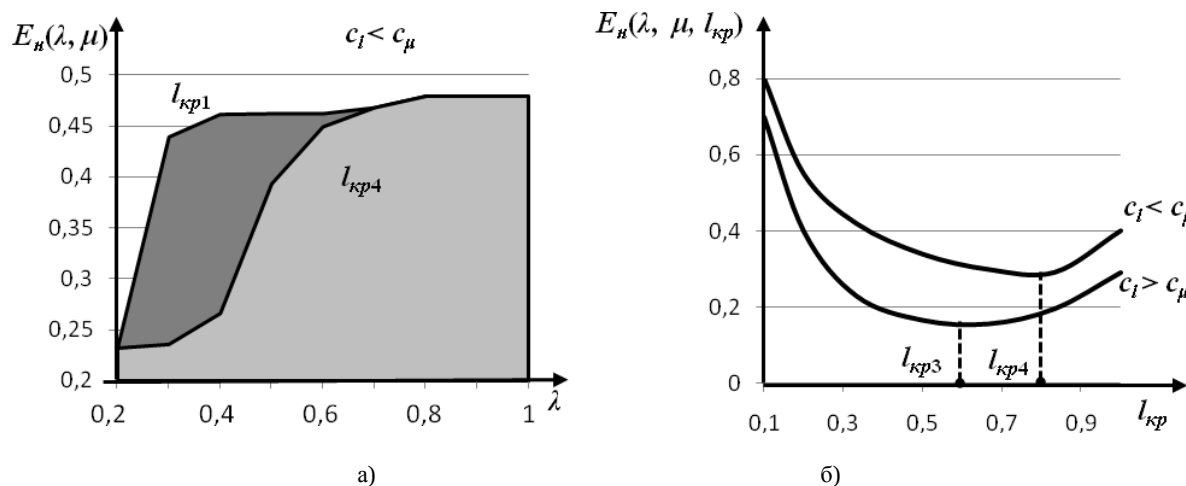


Рисунок 6 – Интегральный критерий эффективности управления вычислительной мощностью ВУ с переменной производительностью: а) зависимость критерия эффективности от интенсивности входного потока заданий; б) зависимость критерия эффективности от критической длины очереди при различных значениях штрафов

Таким образом, управляя значением критической длины очереди l_{kp} , можно изменять производительность ВУ на основании критерия (3).

2. Изучение характера изменения длины средней очереди заданий на обслуживание в условиях переменной мощности ВУ. В ходе модельного эксперимента изменение производительности ВУ осуществлялось скачкообразно в соответствии с выражением (2). Результаты моделирования изменения

мощности ВУ в зависимости от значения критической длины очереди заданий на обслуживание ВУ в среде Simulink показаны на рисунке 5 в виде временных диаграмм.

Временная диаграмма изменения интенсивности входного потока заданий, рассчитанная блоком Repeating Sequence Interpolated, изображена на рисунке 7, а. Можно убедиться, что характер входного потока является пилообразным и соответствует заданному, изображенному на рисунке 3.

В результате моделирования получены следующие характеристики ВУ: коэффициенты использования основного, дополнительного и совместного использования вычислительных каналов, а также средняя длина очереди заданий, ожидающих обслуживания. В данном вычислительном эксперименте определенный интерес представляет собой процесс включения и выключения дополнительного вычислительного канала.

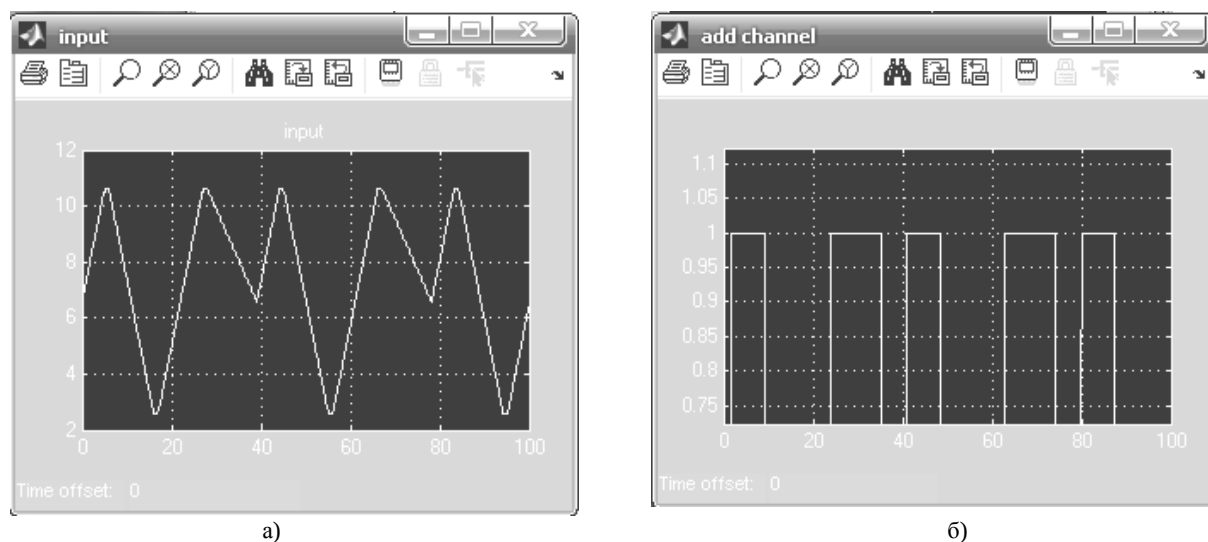


Рисунок 7 – Временные диаграммы результатов моделирования ВУ переменной мощности: а) диаграмма интенсивности входного потока заданий; б) диаграмма подключения дополнительного вычислительного канала

В периоды времени, когда интенсивность входного потока достигает пиковых значений $\lambda(t) \geq \lambda_{кр}$, длина очереди заданий, ожидающих обслуживания, становится также больше критической. В такой ситуации осуществляется подключение дополнительного вычислительного канала, что отображено на диаграмме 7, б. При снижении нагрузки на ВУ, а, следовательно, и снижении длины очереди необходимость в использовании дополнительного канала отпадает, он отключается, и ВУ работает в обычном режиме.

Характер изменения средней длины очереди заданий, ожидающих обслуживания, в условиях переменной мощности ВУ показан на рисунке 8. Для большей наглядности здесь отображен фрагмент временного интервала моделирования.

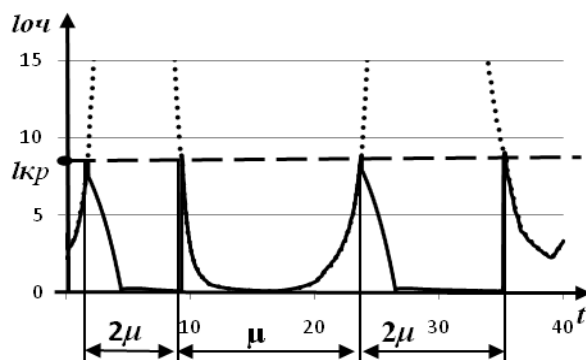


Рисунок 8 – Средняя длина очереди заданий на входе ВУ переменной мощности

установленного критического значения.

Предложенная стратегия управления мощностью ВУ в условиях критической нагрузки позволяет управлять введением в вычислительный процесс дополнительного вычислительного канала при критическом уровне длины очереди заданий на входе узла. Результаты вычислительных экспериментов с моделью подтвердили сформулированное правило управления мощностью ВУ. Представленная модель

Пунктирной линией отмечены значения

длины очереди заданий при использовании единственного вычислительного канала. При пиковых нагрузках длина очереди в такой ситуации может вырасти на порядки, что приведет к потере данных и невыполнению требования непрерывности ИТ-сервиса. При включении дополнительного канала средняя длина очереди заданий начинает уменьшаться, не превышая критического значения $l_{кр}$. При снижении нагрузки на ВУ дополнительный канал отключается, что приводит к увеличению средней длины очереди, но при этом средняя длина очереди заданий, ожидающих обслуживания в ВУ не превышает

ВУ переменной мощности может быть использована в коммуникационных сетях с целью повышения эффективности их функционирования в условиях пиковых нагрузок.

В качестве перспективного направления в дальнейших исследованиях целесообразно рассмотреть модель управления мощностью ВУ в условиях критических нагрузок для произвольного числа подключаемых вычислительных каналов и различном характере изменения входного потока заданий.

Библиографический список использованной литературы

1. Бон Ян Ван. ИТ Сервис-менеджмент, введение // Я.В. Бон, Г. Кеммерлинг, Д. Пондаман; под ред. М.Ю. Потоцкого (русская версия). — М.: IT Expert, 2003. — 215 с.
2. Соколов Ю.Н. Компьютерный анализ и проектирование систем управления: учеб. пособие / Ю.Н. Соколов. — Харьков: Нац. аэрокосм. ун-т «Харьк. авиац. ин-т», 2010. — Ч. 5: Нелинейные системы. — 344 с.
3. Вентцель Е.С. Теория вероятностей // Е.С. Вентцель, Л.А. Овчаров. — М.: Наука, 1973. — 368 с.

Поступила в редакцию 24.02.2012 г.

Скатков О.В., Балакірева І.А. Управління обчислювальною потужністю вузла під час критичних подій

Розглянуто завдання управління потужністю обчислювального вузла в умовах критичного навантаження. Аналізується стратегія ухвалення рішень про введення в роботу додаткового обчислювального каналу в разі завантаження вузла вище критичного. Ефективність запропонованого підходу показана на підставі результатів обчислювального експерименту.

Ключові слова: ІТ-сервіс, обчислювальний вузол, критичне робоче навантаження, стратегія управління, інцидент.

Skatkov A.V., Balakireva I.A. Computational node power control at critical events

The problem of computational node power control at critical working load has been examined. The decision-making strategic of engaging additional computing channel has been analyzed at critical node working load situation. Efficiency of proposing approach is discussed and explained using computer experiment results.

Keywords: IT-service, computational node, critical working load, management strategic, incident.