Лекция 3. Производительность системы функциональных устройств.

1. Оценка реальной производительности ВС.

Любая ВС – совокупность функциональных устройств (ФУ). ФУ бывают двух типов:
Простое ФУ – если на нем никакая другая операция не может начаться, пока не кончится предыдущая (монопольное использование всего оборудования для выполнения операции).
Конвеерное ФУ – ФУ, которое может выполнять одновременно несколько элементарных операций в режиме конвеера (обычно это линейная цепочка простых элементарных ФУ, имеющих одинаковые времена срабатывания). Элементарные ФУ – ***ступени конвеера***, их число – ***длина конвеера***. Простое ФУ = конвеерное ФУ с длиной 1. Все ФУ выполняют одну или несколько операций, скорость их работы измеряется в операциях в секунду.
Предположение 1 – время выполнения конкретной операции на конкретном ФУ фиксировано, не равно 0 и не зависит от других ФУ.
Предположение 2 – любое ФУ не может одновременно выполнять операцию и сохранять результат предыдущего срабатывания – т.е. не имеет собственной памяти. Результат предыдущего срабатывания хранится в нем до начала следующей операции, а затем пропадает.
Предположение 3 – все ФУ работают по индивидуальным командам: команда + аргументы попадают на вход ФУ. При этом не происходит тупиковых ситуаций.
*Предположение 4 – Если отрезок времени работы ФУ равен **T**, а длительность операции **τ**, то целое число операций **Q**, выполненных за время **T,** равно либо **[T/τ]**, либо **[T/τ]-1**. При больших **T** эти величины асимптотически совпадают.*
Стоимость операции ФУ – время ее реализации на ФУ (время срабатывания ФУ).
Стоимость работы ФУ – время последовательной реализации всех операций на простых ФУ с аналогичными временами срабатывания.
Загруженность ФУ (***0<=Z<=1***) – отношение стоимости реально выполненной работы S_R за фиксированный промежуток времени ***T*** к максимально возможной стоимости S_M. ***Q*** – число выполненных ФУ операций за время ***T***.

Утв. 1. $S_M = T$ ***для простого ФУ,*** $S_M = L*T$ ***для конвеерного ФУ с длиной конвеера L,*** $S_R = Q*\tau$.

Реальная производительность ВС (P_R) – количество операций, выполненных ВС в среднем за единицу времени.
Пиковая производительность ВС (P_M) – максимальное количество операций, которое может быть выполнено ВС за единицу времени при отсутствии связей между ее ФУ.

Утв. 2. $P_R = \sum_{k=1}^{m} Z_k P_{M,k}$, $P_M = \sum_{k=1}^{m} P_{M,k}$, где m – кол-во ФУ, $P_{M,k}$ и Z_k – их пиковые производительности и загруженности. Док-во:
Простое ФУ (L=1): $P_M=1/\tau$, $S_R=Q*\tau$, $S_M=T$, $Z=S_R/S_M=Q*\tau/T$, $P_R = Q/T = (Q*\tau/T) * (1/\tau) = Z * P_M$ – ч.т.д.
Конвеерное ФУ (L>1): $P_M=L/\tau$, $S_R=Q*\tau$, $S_M=L*T$, $Z=(Q*\tau)/(L*T)$, $P_R=Q/T=(Q*\tau)/(L*T) * (L/\tau) = Z * P_M$ – ч.т.д.
Аддитивность формул доказывает Утв. 2 для любого кол-ва устройств.
Для одного устройства: $P_R = Z * P_M$ => для повышения P_R надо увеличивать Z.

Для системы устройств: $Z = \sum_{k=1}^{m} \alpha_k Z_k$, $\alpha_k = P_{M,k} / P_M \geq 0$, $\sum_{k=1}^{m} \alpha_k = 1$ => 0<= Z <= 1.

Вывод: для увеличения загруженности системы необходимо повышать загруженность каждого ФУ.
Пример: ВС имеет два устройства - сумматор и умножитель с одинаковой пиковой производительностью. Решаем задачу C = A + B. При обычной схеме решения имеем $P_R = 0.5 P_M$. Можно ли повысить реальную производительность? Да: C = A + 1*B => $P_R = P_M$ – но кому это нужно?

2. Ускорение и эффективность использования МВС.

Опр. 1. Ускорение - отношение $S_m = T_1/T_m$ ($0<=S_m<=m$) времен выполнения одной задачи на системе из m и одного одинаковых процессоров (ФУ). ***Эффективность*** $E_m = S_m /m$ или $E_m = S_m /m*100\%$.
Это определение годится только для систем с одинаковыми процессорами (ФУ). На неоднородных ВС с различными ФУ ***эффективность совпадает с загруженностью***, а ускорение определяется иначе:

Опр. 2. Ускорение на неоднородной ВС есть $S_m = P_R / \max_{1\leq k\leq m} P_{M,k} = \sum_{k=1}^{m} Z_k P_{M,k} / \max_{1\leq k\leq m} P_{M,k}$.

Утв. 3. В системе из m устройств с одинаковой пиковой производительностью ($P_{M,1} = \ldots = P_{M,m}$***):***

1) $Z = \frac{1}{m}\sum_{k=1}^{m} Z_k$ 2) $P_R = \sum_{k=1}^{m} P_{R,k}$ 3) $P_M = mP_{M,1}$ 4) $S_m = \sum_{k=1}^{m} Z_k$ 5) если все ФУ простые => $S_m = T_1 / T_m$.

Рассмотрим ВС из m простых ФУ. Пусть между ними установлены какие-то функциональные связи, которые не меняются со временем. Их можно отобразить в виде ориентированного мультиграфа: точки – ФУ, дуги – связи между ФУ (если результат с выхода одного ФУ передается сразу на вход другого). Назовем этот мультиграф ***графом ВС***. Предположим: исходные данные введены в ВС мгновенно, результаты работы одного ФУ передаются другому без задержки. Исследуем ***максимальную производительность*** такой ВС.

Утв. 4. Если все ФУ простые, а граф ВС связный => $P_{R,\max} = m \cdot \min_{1 \le k \le m} P_{M,k}$.

Следствие 1: При выполнении условий Утв. 4:
1) асимптотически каждое ФУ выполняет одно и тоже число операций в секунду;
2) загруженность любого ФУ не превосходит загруженность самого непроизводительного ФУ;
3) если загруженность какого-либо ФУ = 1, то это самое непроизводительное ФУ;
4) загруженность ВС $Z \le Z_{\max} = m \cdot \min_{1 \le k \le m} P_{M,k} / P_M$;
5) ускорение ВС $S_m \le S_{m,\max}^{(tech)} = m \cdot \min_{1 \le k \le m} P_{M,k} / \max_{1 \le k \le m} P_{M,k}$.

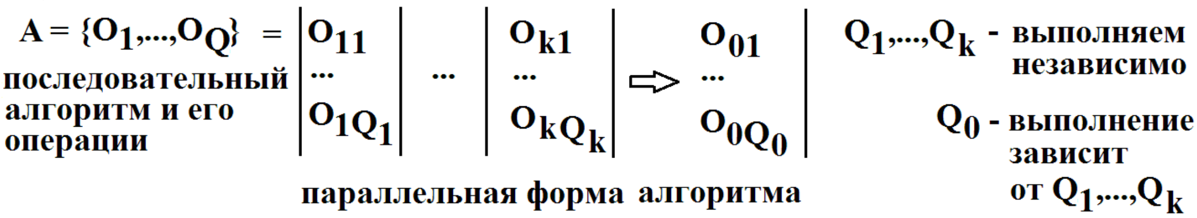
Следствие 2 (1-ый Закон Амдала): Производительность ВС, состоящей из связанных ФУ, в общем случае определяется самым непроизводительным устройством.

Следствие 3: Если все ФУ простые, а граф ВС связный => асимптотическая производительность ВС будет максимальной, если все устройства имеют одинаковые пиковые производительности.

Смысл: Расписание подачи команд должно минимизировать простои устройств!

3. Связь аппаратных и программных ограничений.

Рассмотрим ВС из m простых универсальных ФУ. Пусть Q – число всех элементарных операций алгоритма, q – максимальное число элементарных операций в параллельных ветвях алгоритма (***высота параллельной формы алгоритма***), k – число параллельных ветвей (***ширина параллельной формы алгоритма***).



Утв. 5. Максимально возможное ускорение $S_m \le S_{m,\max}^{(\mathrm{alg})} = Q/q$ при любом числе устройств m.

<u>Док-во:</u> из утв. 3-4) имеем $S_m = Z_1 + \ldots + Z_m = Q_1 * \tau/T + \ldots + Q_m * \tau/T = Q * \tau/T$, $T >= q * \tau \Rightarrow S_m <= Q/q$.

Следствие 1: Минимальное число ФУ ВС, при котором может быть достигнуто максимально возможное ускорение, равно ширине алгоритма, т.е. m >= k.

Пусть Q_0 операций из Q выполняются последовательно и $\beta = Q_0/Q$ – ***доля последовательных вычислений***.

Следствие 2 (2-й закон Амдала): Если ВС состоит из m одинаковых простых универсальных ФУ и при выполнении параллельной части алгоритма они все загружены полностью, то $S_{m,\max} = m / [m*\beta+(1-\beta)]$.

Док-во: из утв. 3-4) $S_m = Z_1 + \ldots + Z_m$, и все последовательные операции выполняются на первом ФУ => $Z_1 = [\beta*Q*\tau+ (1-\beta)*Q*\tau/m]/T=1$, $Z_k = [(1-\beta)*Q*\tau/m]/T$ $(k>1)$ => $S_m = 1 + (m-1)*[(1-\beta)*Q*\tau/m]/[\beta*Q*\tau+ (1-\beta)*Q*\tau/m]$ => $S_m = 1 + (m-1)*(1-\beta)/[m*\beta+(1-\beta)] = m / [m*\beta+(1-\beta)]$ – ч.т.д.

Следствие 3 (3-й закон Амдала): Если ВС состоит из простых одинаковых универсальных ФУ, то при любом режиме работы ее ускорение <= 1/β.

Зам. 1: На МВС с большим числом процессоров эффективно решаются задачи, где β=0.01-0.001.

Зам. 2: 3-й закон Амдала используется для прогнозирования ускорения.

Резюме: Из характеристик ВС получаем m и $S_{m,\max}^{(tech)}$, из свойств алгоритма k и $S_{k,\max}^{(\mathrm{alg})}$ и сравниваем. Первичен в любом случае алгоритм, то есть используемое в реальных вычислениях $m \le k$.