Técnicas Básicas

- **Aferição**
  - **Avaliação experimental, Medição ou Prototipação**
    - Resultados precisos
    - Custo alto
    - Não permite generalização
- **Modelagem**
  - **Simulação**
    - Precisão média
    - Custo médio
    - Generalização moderada
  - **Analítica**
    - Precisão relativa
    - Baixo custo
    - Permite generalização

Aferição

- Índices de Desempenho em Programas
  - Benchmarks

- CPU
  - Tempo de execução
  - Taxa de utilização
  - Ociosidade
- Memória
  - Ocupação
  - Swap
  - Cache
- Rede
  - Número de comunicações
  - Volume de comunicações

Aferição

O que avaliar no contexto da Programação Paralela?

- Ganho de Desempenho na execução de um programa:
  - Tempo de execução
  - Utilização do processador
  - Consumo energético
  - Consumo de memória
  - Acessos ao cache
  - Volume e número de comunicações

Aferição

O que avaliar no contexto da Programação Paralela?

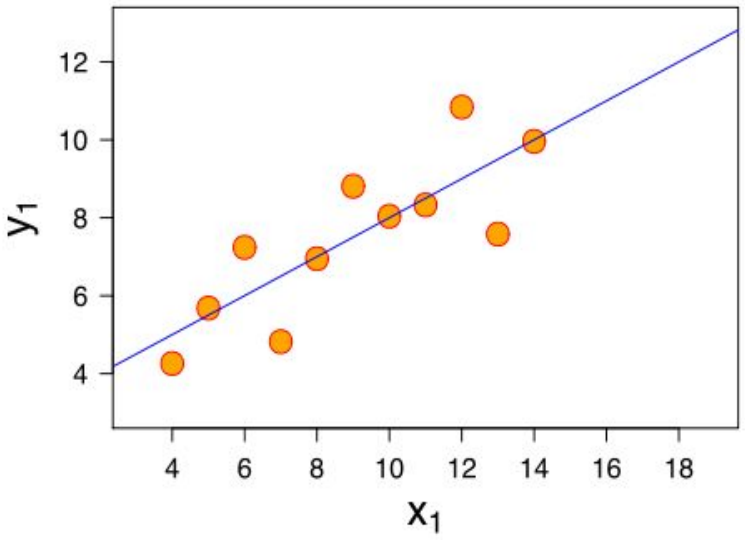
- Cuidar como avaliar.
  - A avaliação *clássica* considera o tempo de execução, sendo o ganho de desempenho avaliado comparando o tempo de execução obtido com a aplicação de diferentes estratégias/técnicas de implementação
- Caso típico: Tempo de execução
  - Várias (30, pelo menos) execuções
  - Máquina dedicada, com warm up realizado
  - Cuidar otimizações agressivas

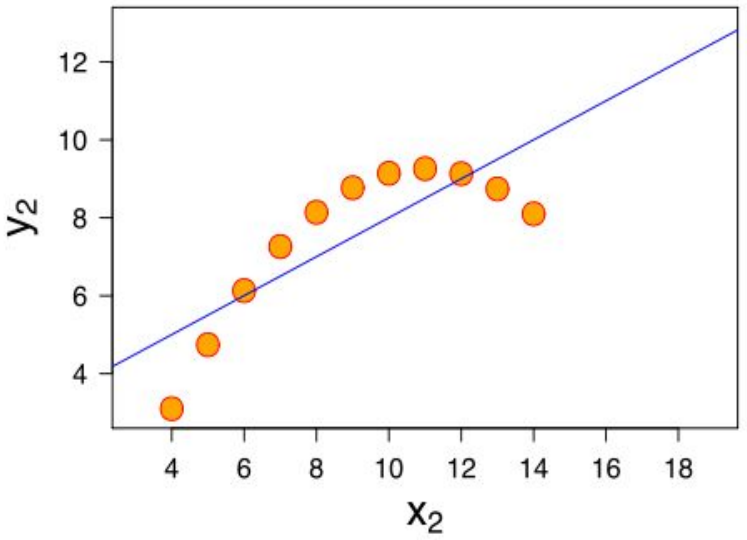
- Métricas descritivas
  - Um ponto de execução não é significativo, então, quantos experimentos realizar?
  - Uma média, apresentada com seu desvio padrão, tem maior representatividade
  - Média e desvio padrão descrevem uma coleção de dados representando uma amostra

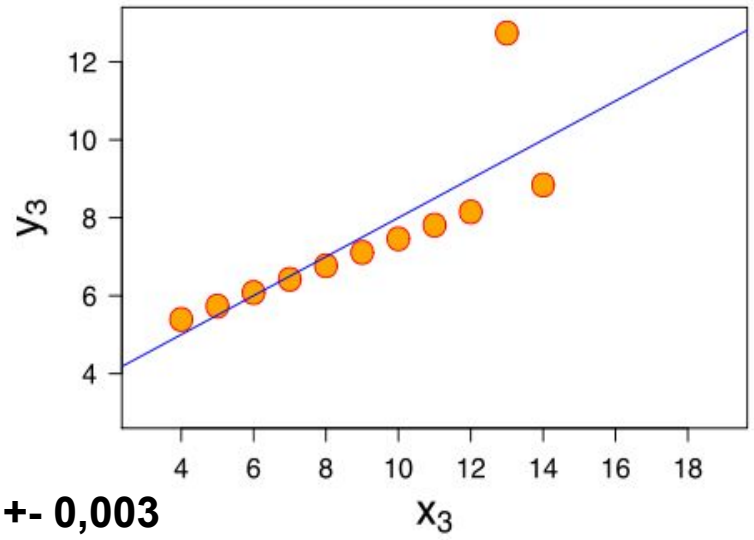
Cuidado: “Quarteto de Anscombe”

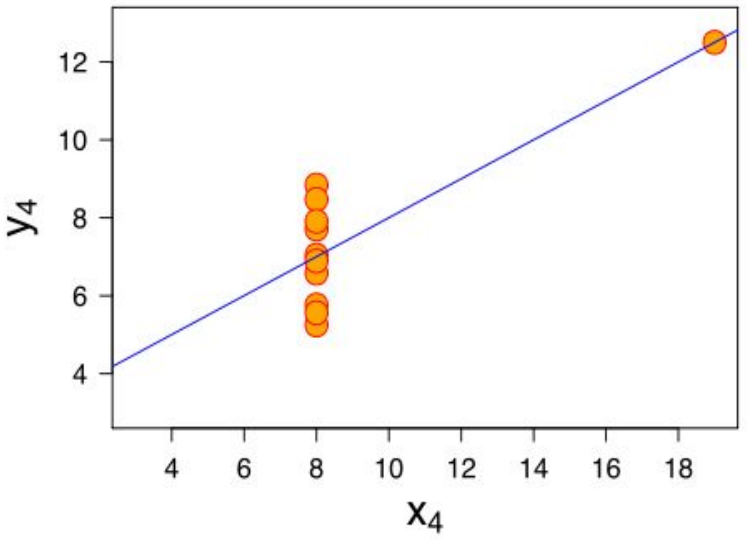
Quarteto de Anscombe

Quarteto de Anscombe é o nome dado a quatro conjuntos de dados que aparentam ser idênticos quando descritos por certas técnicas de estatística descritiva (como a média e a variância), mas que são muito distintos quando exibidos graficamente. Ele leva o nome do estatístico F.J. Anscombe que o publicou pela primeira vez em 1973, com o objetivo de demonstrar tanto a importância de se visualizar os dados antes de analisá-los quanto o efeito dos outliers nas propriedades estatísticas. (wikipedia)





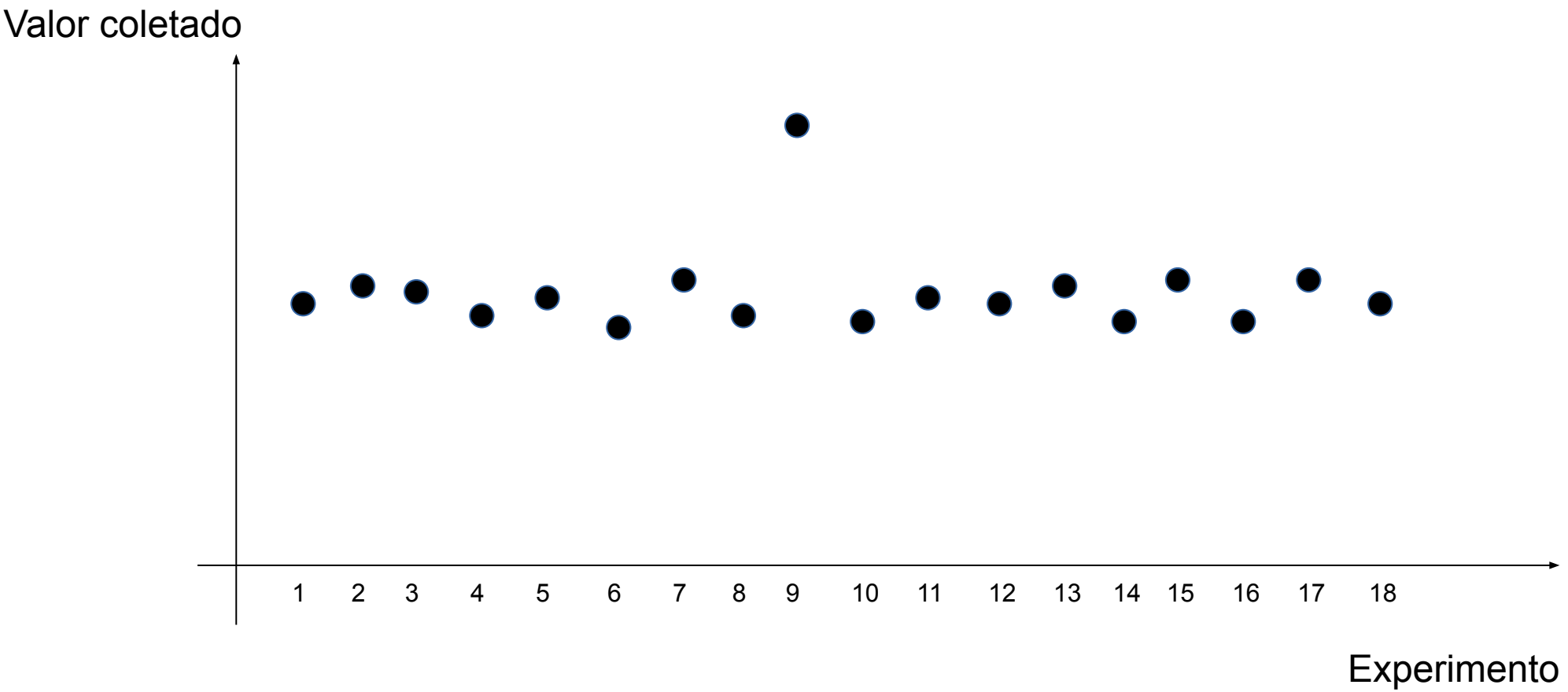




Média de x: 9
Variância de x: 11
Média de y: 7,50
Variância de y: 4,12 +- 0,003

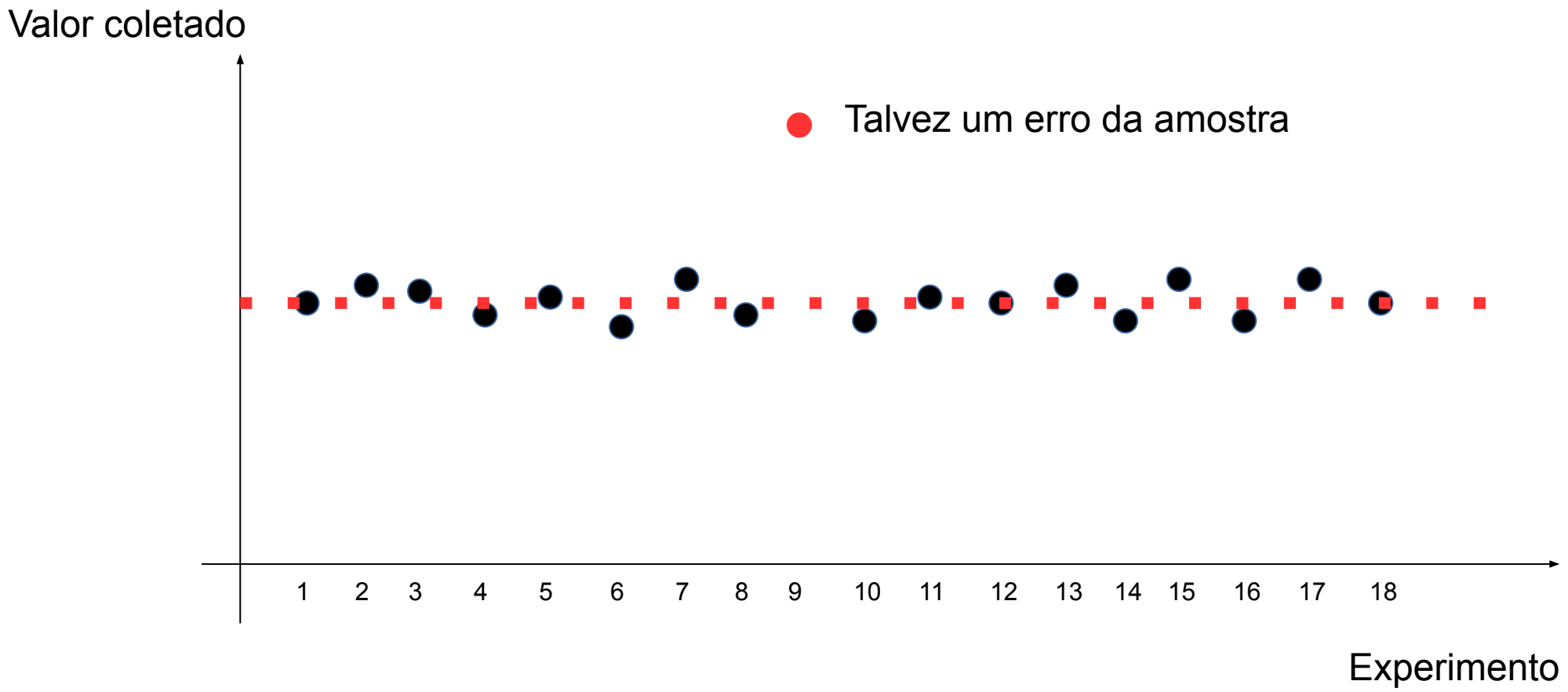
Apresentação dos Dados

- Nuvem de Dispersão
  - Visualização dos experimentos



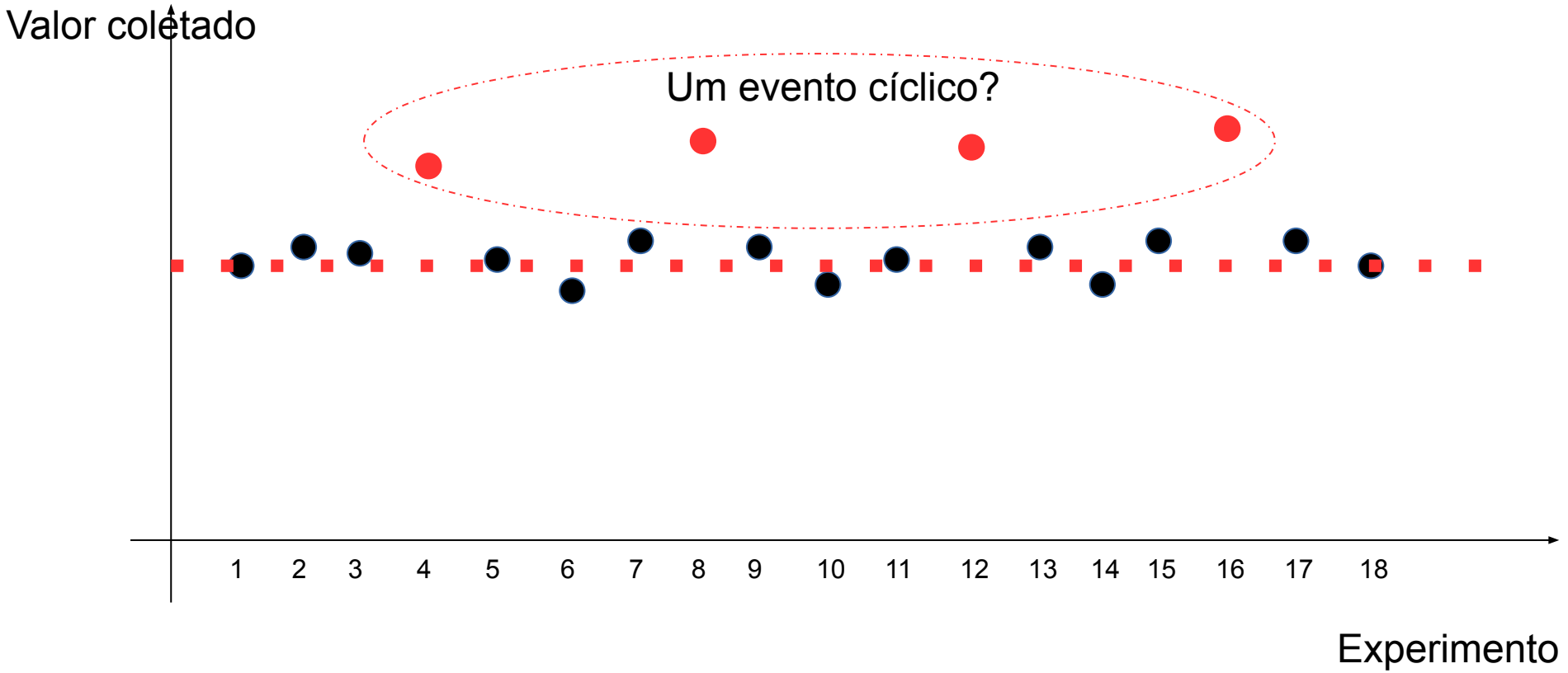
Apresentação dos Dados

- Nuvem de Dispersão
  - Visualização dos experimentos



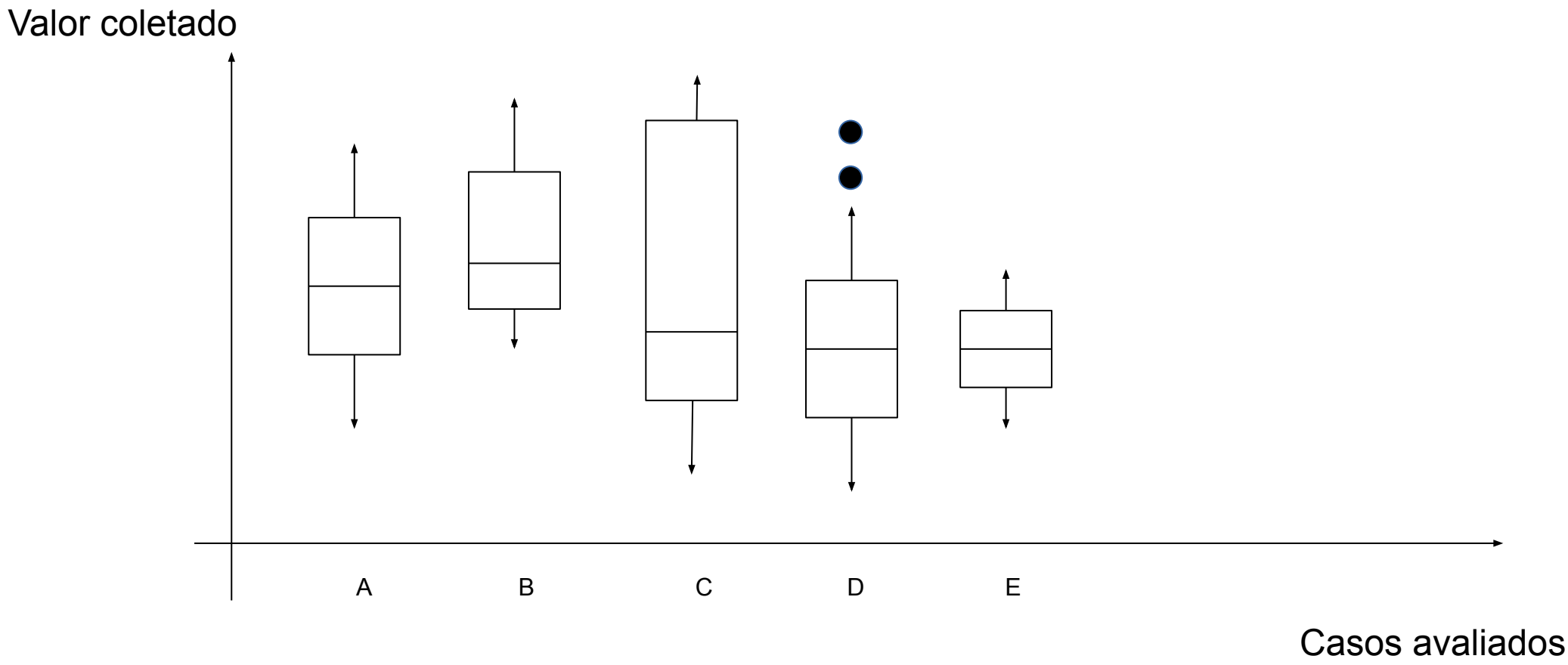
Apresentação dos Dados

- Nuvem de Dispersão
  - Visualização dos experimentos



Apresentação dos Dados

- Gráfico de Caixa (Boxplot)
  - Visualização dos experimentos



Apresentação dos Dados

- Testes de Hipóteses
  - Decidir se uma conjectura sobre determinada característica de uma ou mais populações é, ou não, apoiada pela evidência obtida dos dados amostrais.
  - Conjectura == hipótese estatística
  - Deve-se decidir se ela é verdadeira ou não com base nos dados coletados.
  - Os testes de hipóteses permitem verificar se existe uma diferença real (significativa) ou aleatória no processo em análise.

Métricas

Makespan: Tempo de Execução

Corresponde ao tempo necessário para executar uma tarefa. Em ambientes paralelos:

$$T_S \neq T_1$$

$$T_1 = T_S + \sigma$$

Onde:

- ***p*** é o número de unidades ativas (processadores) utilizadas
- **σ** é o sobrecusto da gestão do paralelismo
- T_S *é o tempo de execução da aplicação na versão sequencial*
- T_p *é o tempo de execução na versão paralela*
- T_1 *é o tempo de execução na versão paralela com um processador*

Métricas

SpeedUp: Fator de aceleração

Indica quantas vezes o uma alternativa do programa é mais rápida que a versão original para executar uma dada tarefa. Consiste na razão entre o tempo obtido pela versão original e a nova proposta.

Muito usado para avaliação de ambientes paralelos. É determinado pela razão entre o melhor tempo sequencial e o melhor tempo da versão paralela

SpeedUp(p) **ou** $S_p = T_S / T_p$

Onde:

p é o número de unidades ativas (processadores) utilizadas

T_s *é o tempo de execução da aplicação na versão sequencial*

T_p *é o tempo de execução na versão paralela*

Se $S_p > 1$ a versão paralela reduziu o tempo de execução (ficou mais rápido que a sequencial)

Se $S_p < 1$ a versão paralela aumentou o tempo de execução (ficou mais lenta que a sequencial)

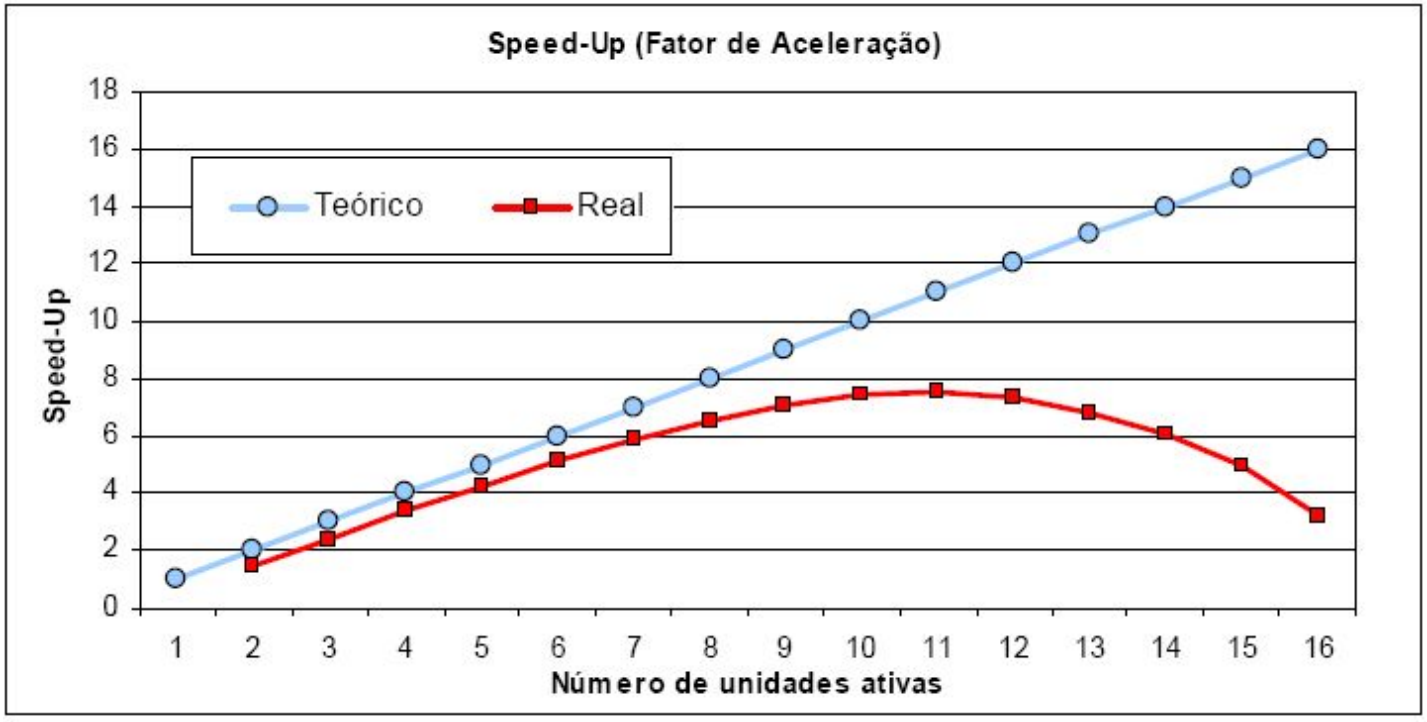
Métricas

SpeedUp: Fator de aceleração

Cada aplicação tem sua curva que depende do trabalho e da incidência de operações relacionadas à gestão da execução (paralela)

Todo o algoritmo de uma aplicação tem um número de unidades ativas ideal para a obtenção do melhor desempenho em uma dada arquitetura alvo

Não é verdade que quanto mais unidades ativas melhor - aumenta o sobrecusto de gestão e diminui a eficiência



Métricas

SpeedUp: Fator de aceleração

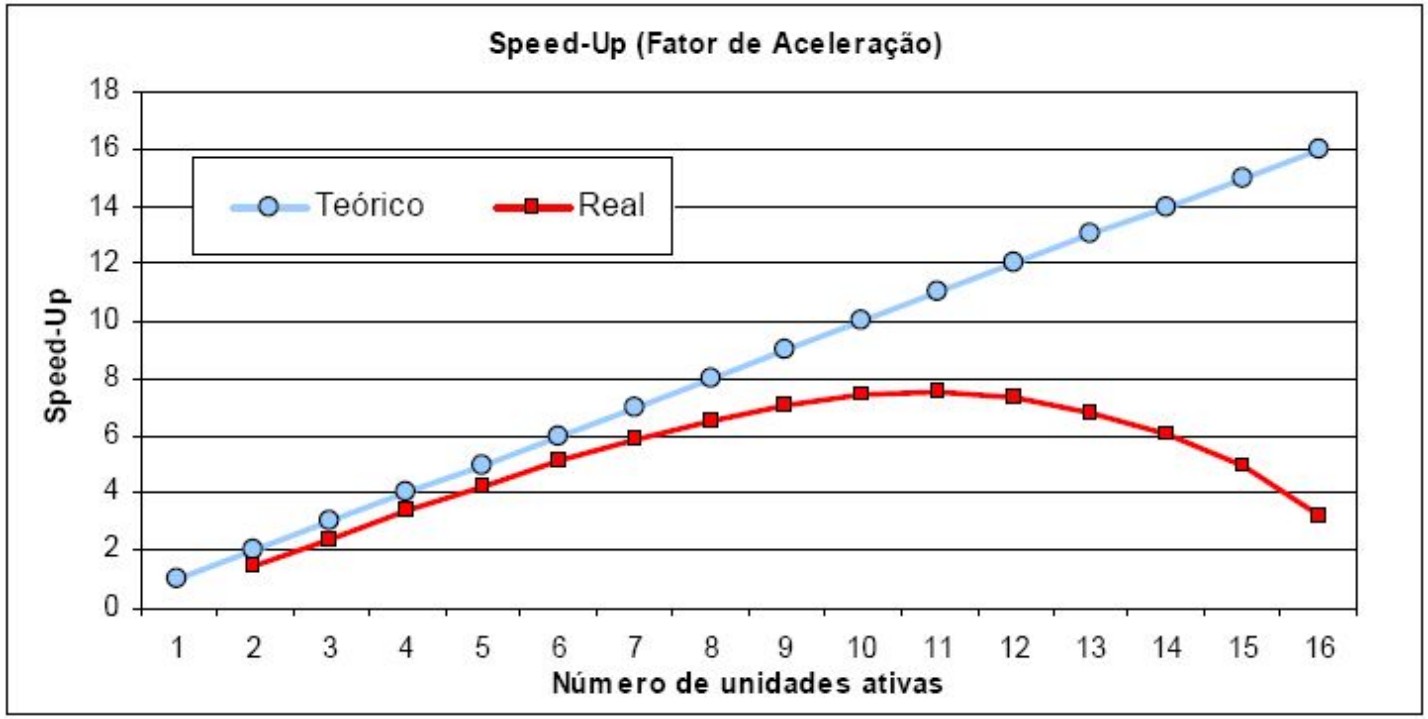
Cada aplicação tem sua curva que depende do trabalho e da incidência de operações relacionadas à gestão da execução (paralela)

Todo o algoritmo de uma aplicação tem um número de unidades ativas ideal para a obtenção do melhor desempenho em uma dada arquitetura alvo

Não é verdade que quanto mais unidades ativas melhor - aumenta o sobrecusto de gestão e diminui a eficiência

Note que, usualmente:

$$T_S / T_p < T_1 / T_p$$



Métricas

Eficiência: Aproveitamento dos recursos

Muito usado em paralelismo, indica a taxa de utilização média das unidades ativas

É calculado pela razão entre o *SpeedUp* e o número de *unidades ativas* utilizadas

Eficiência(p) ou $E_p = S_p / p$

Normalmente, as unidades ativas ficam parte de seu tempo esperando por resultados de vizinhos

Reduz sua taxa de utilização e consequentemente a eficiência

Métricas

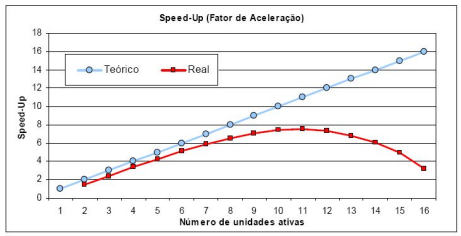
Eficiência: Aproveitamento dos recursos

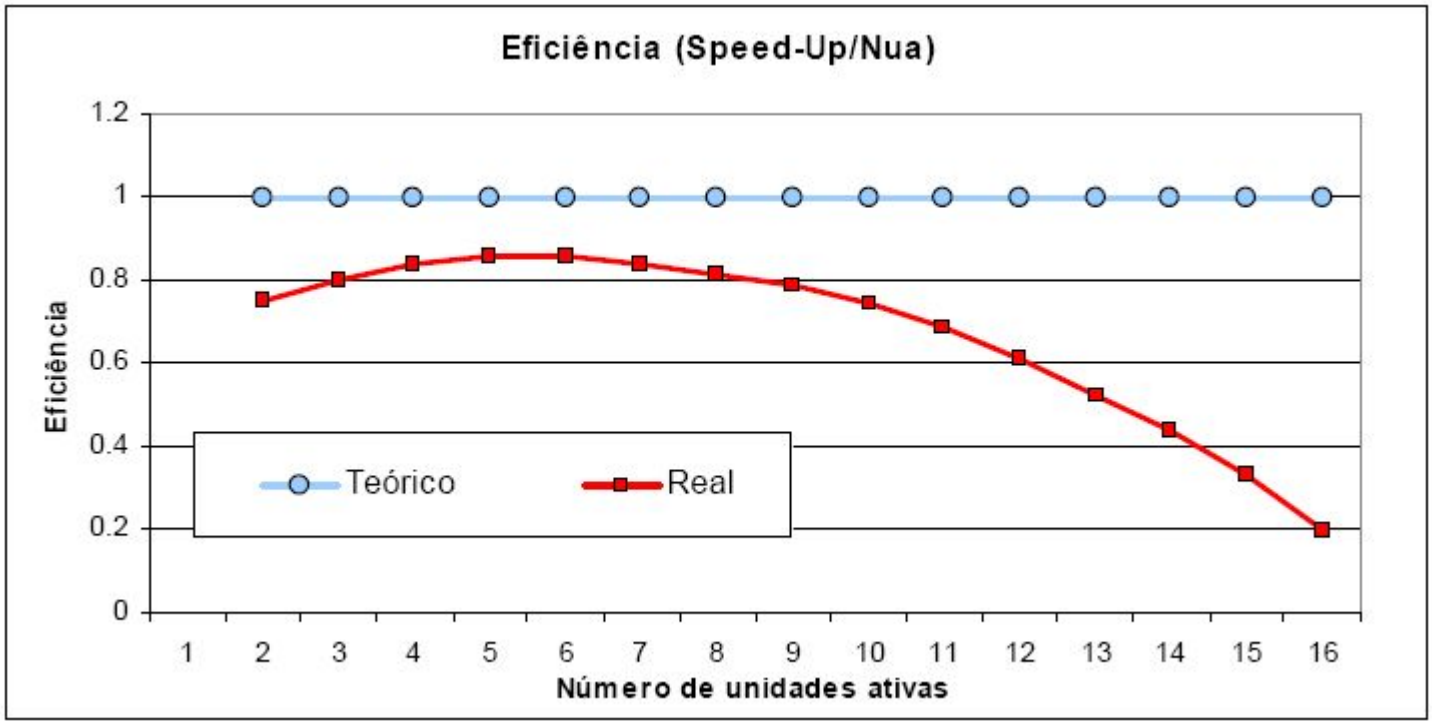
Eficiência ideal

Cada unidade ativa com 100% do tempo ativa (linha azul)

A melhor taxa de utilização média não significa o menor tempo de execução

Exemplo: o menor tempo de execução ocorreu com 11 unidades ativas e a melhor taxa de utilização média com 5 unidades ativas





Métricas

Latência: Custo da comunicação

Tempo necessário para enviar mensagem através da rede de interconexão

Inclui tempo de empacotar e desempacotar dados mais tempo de envio propriamente dito

A latência aumenta à medida que a quantidade de dados a serem enviados aumenta

O aumento não é linear

A componente do tempo referente ao custo de empacotamento e desempacotamento não varia tanto em relação ao tamanho da mensagem como a componente de custo de envio pela rede

Métricas

Vazão: Desempenho da comunicação

Expressa a capacidade da rede de "bombear" dados entre dois pontos

Unidade

Quantidade de dados por unidade de tempo

Ex: 10 MBytes/segundo (10MB/s)

A vazão (V) é afetada pela "largura" (L) do canal de comunicação (bits) e pela frequência (F) da transmissão dos dados (MHz)

$$V \propto L \times F$$

Métricas

Latência e Vazão: Desempenho da comunicação

Exemplos:

Latência de 1 mensagem de 1 byte entre máquinas rodando GNU/Linux ligadas por Fast-Ethernet é de aproximadamente 150 µs

A melhor vazão, obtida com uma mensagem de aproximadamente 64 KB, é em torno de 10 MB/s. Próximo do limite teórico (12,5 MB = 100 Mbits/s)