

Análise de Valores Limites de Desempenho

Profa. Jussara M. Almeida
1º Semestre de 2014

Análise de Valores Limites

- Estratégia baseada em modelos de filas mais simples para análise de desempenho
- Obtenção de limites superior e inferior para throughput e tempo de resposta
 - Por Lei de Little, limites para outras métricas podem ser obtidos
- Duas classes de limites de desempenho
 - Limites assintóticos:
 - Mais geral, mais simples de calcular
 - Limites para sistemas balanceados
 - Limites mais precisos

Características

- Provê insights interessantes nos fatores primários que afetam desempenho do sistema
 - Influência do gargalo / bottleneck do sistema
- Adequado para um primeiro corte, para eliminar alternativas no início do estudo
 - Cálculo rápido
 - Planejamento de capacidade de sistemas
- Útil para estimar potencial ganho de desempenho de upgrades alternativos no sistema.
- Tipicamente aplicado para o caso de uma única classe (existem generalizações para múltiplas classes)

Parâmetros dos Modelos

- K : número de centros de serviços
- D : soma das demandas por serviço em cada centro (demanda total)
- $D_{\max}$: maior demanda por serviço em um centro
- Tipo da classe de clientes: batch, terminal/interativa ou transação
- Z : think time médio (se classe é do tipo interativa)

Tipos dos Limites

- **Limites otimistas:** máximo throughput e mínimo tempo de resposta possíveis para uma dada carga (λ ou N ou N e Z)
- **Limites pessimistas:** mínimo throughput e máximo tempo de resposta possíveis para uma dada carga

Limites Assintóticos

- Derivados considerando condições extremas (assintóticas) de carga leve e de carga pesada
- Premissa de validação:
 - A demanda por serviço de um cliente em um centro *não depende* do número total de clientes correntemente no sistema e nem em quais centros de serviços eles estão localizados

Limites Assintóticos

- Carga de transação : modelo aberto
- Carga de batch: modelo fechado
- Carga interativa/terminal: modelo fechado

Limites Assintóticos para Cargas de Transação (Aberto)

- Limite superior para o throughput:
 - Indica a máxima taxa de chegada λ que pode ser processada pelo sistema com sucesso
 - Caso λ exceda este valor: tempo de espera na fila é arbitrariamente longo; sistema está saturado
- Estratégia de derivação: aplicar Lei da Utilização

Limites Assintóticos para Cargas de Transação (Aberto)

Limite superior para throughput

Para cada centro k : $U_k = X_k S_k = X V_k S_k = \lambda D_k$

$$U_{\max}(\lambda) = \lambda D_{\max} \leq 1$$

$$\lambda_{\text{sat}} = 1/D_{\max}$$

Sistema saturado : $\lambda > \lambda_{\text{sat}}$

Sistema capaz de processar carga : $\lambda \leq \lambda_{\text{sat}}$

Limites Assintóticos para Cargas de Transação (Aberto)

Limites para tempo de resposta ($\lambda \leq \lambda_{\text{sat}}$)

Limite inferior: não há atrasos na fila

$$R \geq D$$

Limite superior: não há
(chegadas em rajadas: n clientes em n/λ seg)

Limites Assintóticos para Cargas Interativas (Fechado)

Limite superior para throughput

- Primeiro cenário: carga pesada

Para cada centro k : $U_k(N) = X(N) D_k \leq 1$

$$X(N) \leq 1/D_{\max}$$

- Próximo cenário: carga leve, não há enfileiramento

Cada cliente gasta $D + Z$ por interação

$$X(N) \leq N / (D+Z)$$

$$X(N) \leq \min (1/D_{\max} , N / (D+Z))$$

Limites Assintóticos para Cargas Interativas (Fechado)

Limite inferior para throughput

- Cada cliente sempre encontra todos os outros $N-1$ clientes nas filas de todos os centros

Cada cliente gasta $(N-1)D + D + Z$ por interação

$$X(N) \geq N / (ND + Z)$$

$$N / (ND + Z) \leq X(N) \leq \min(1/D_{\max}, N / (D + Z))$$

$$N^* = (D + Z) / D_{\max} : \text{ponto ótimo de operação}$$

Limites Assintóticos para Cargas Interativas (Fechado)

Limites para tempo de resposta
(aplicar Little)

$$N / (ND + Z) \leq X(N) \leq \min(1/D_{\max}, N / (D+Z))$$

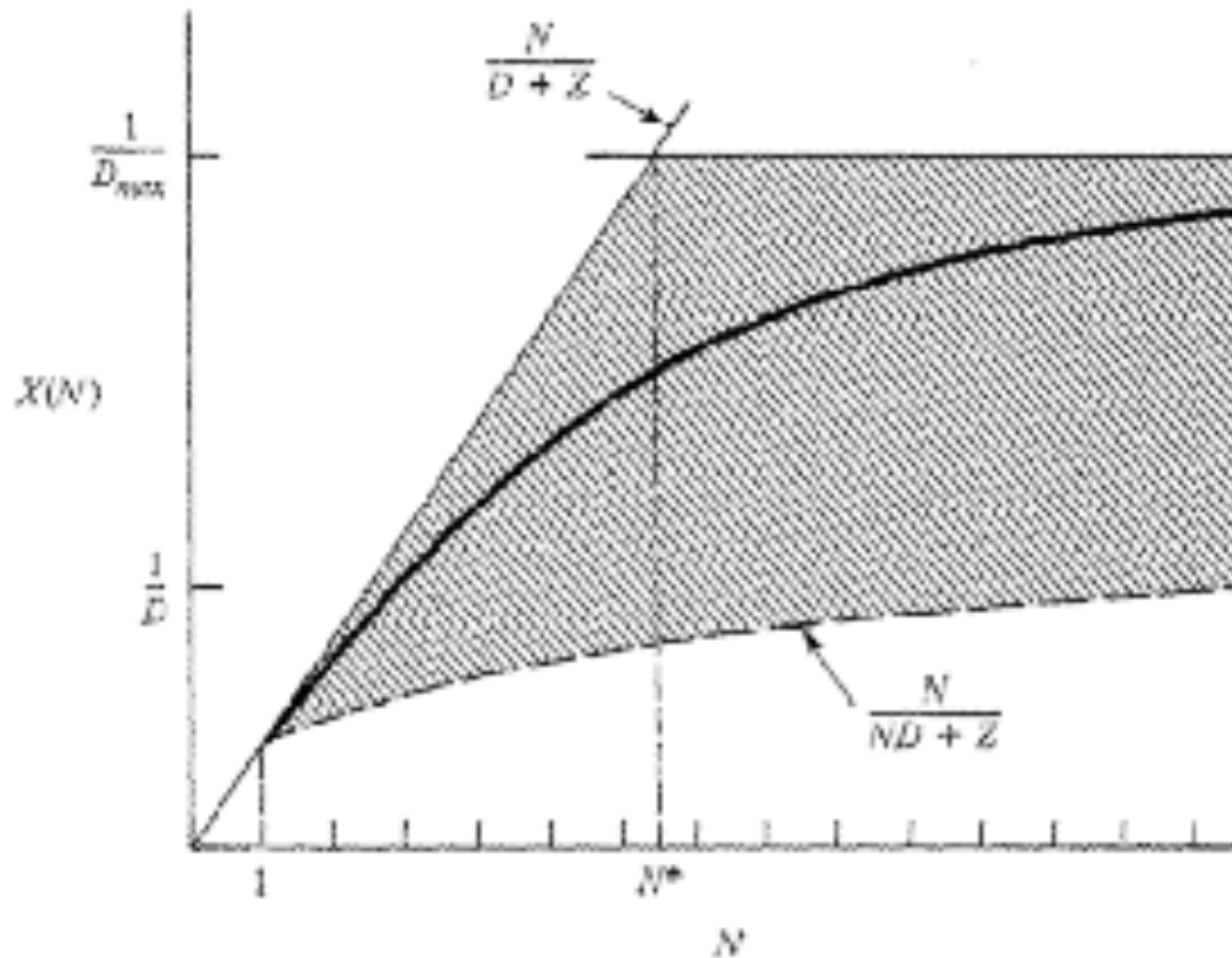
$$\frac{N}{ND+Z} \leq \frac{N}{R(N)+Z} \leq \min\left(\frac{1}{D_{\max}}, \frac{N}{D+Z}\right)$$

$$\max\left(D_{\max}, \frac{D+Z}{N}\right) \leq \frac{R(N)+Z}{N} \leq \frac{ND+Z}{N}$$

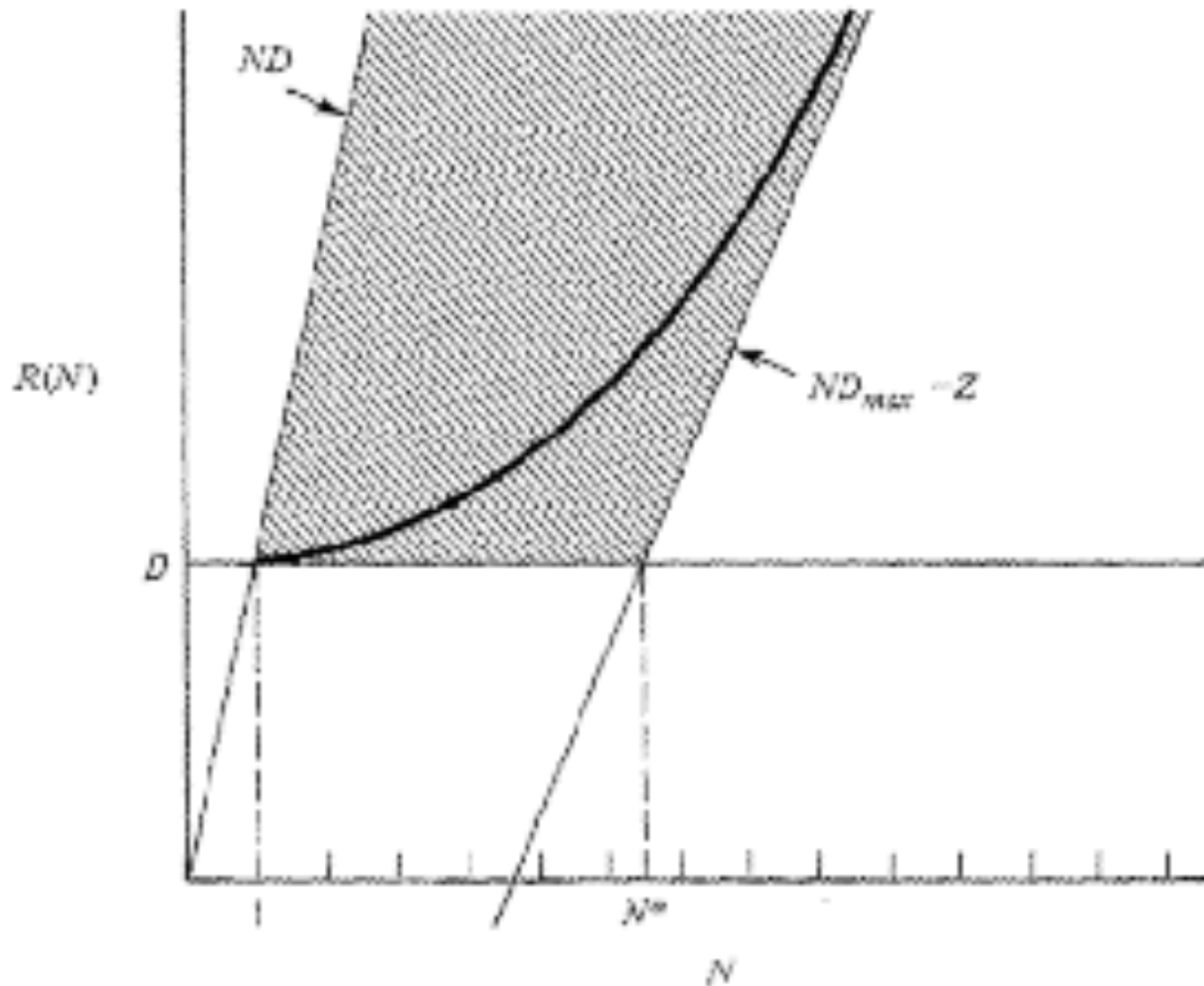
$$\max(D, ND_{\max} - Z) \leq R(N) \leq ND$$

$$N^* = (D + Z) / D_{\max} : \text{ponto ótimo de operação}$$

Limites Assintóticos para Cargas Interativas (Fechado)



Limites Assintóticos para Cargas Interativas (Fechado)



Limites Assintóticos para Cargas Batch (Fechado)

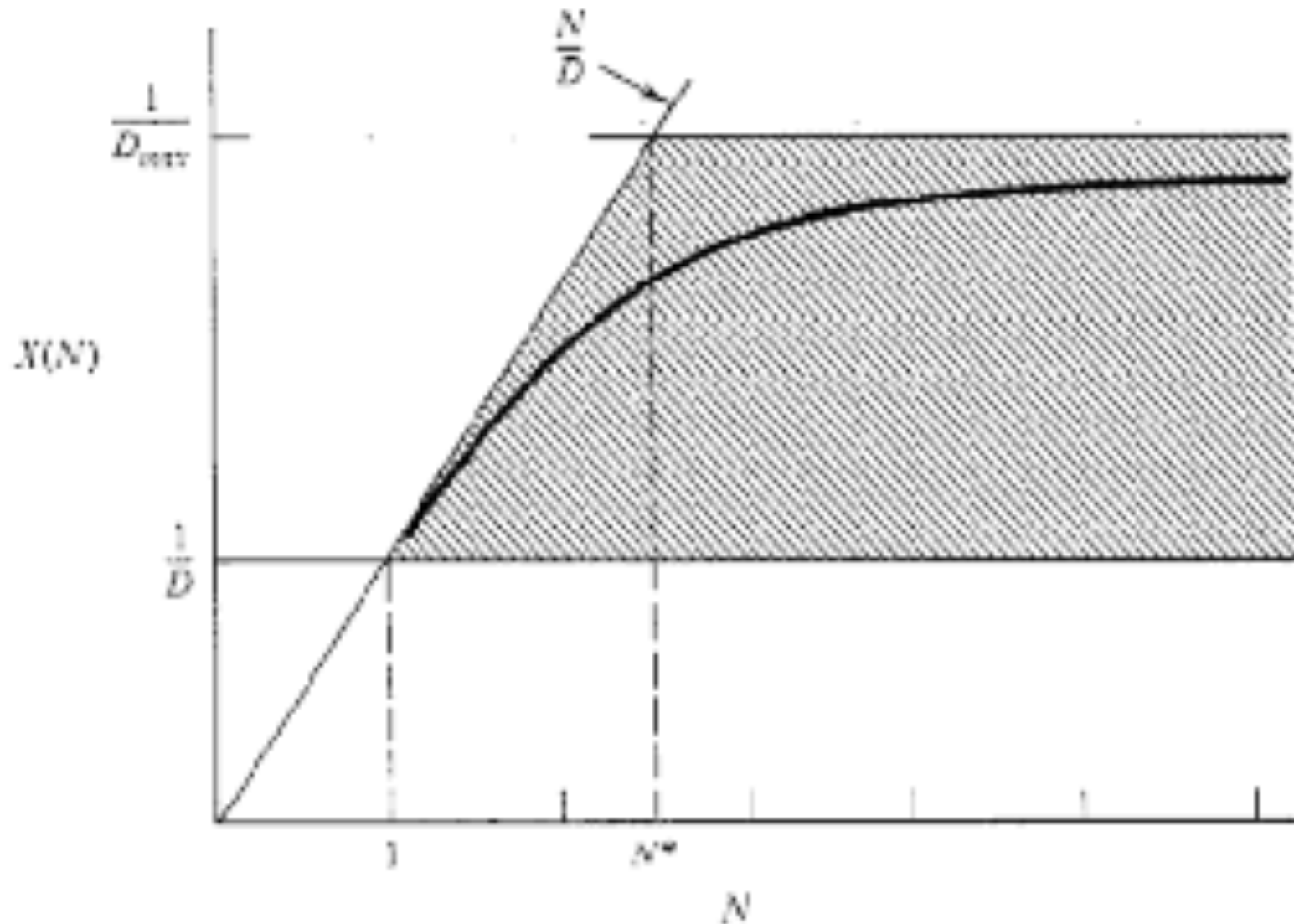
- Utilizar limites para cargas interativas com $Z = 0$

$$1/D \leq X(N) \leq \min(1/D_{\max}, N/D)$$

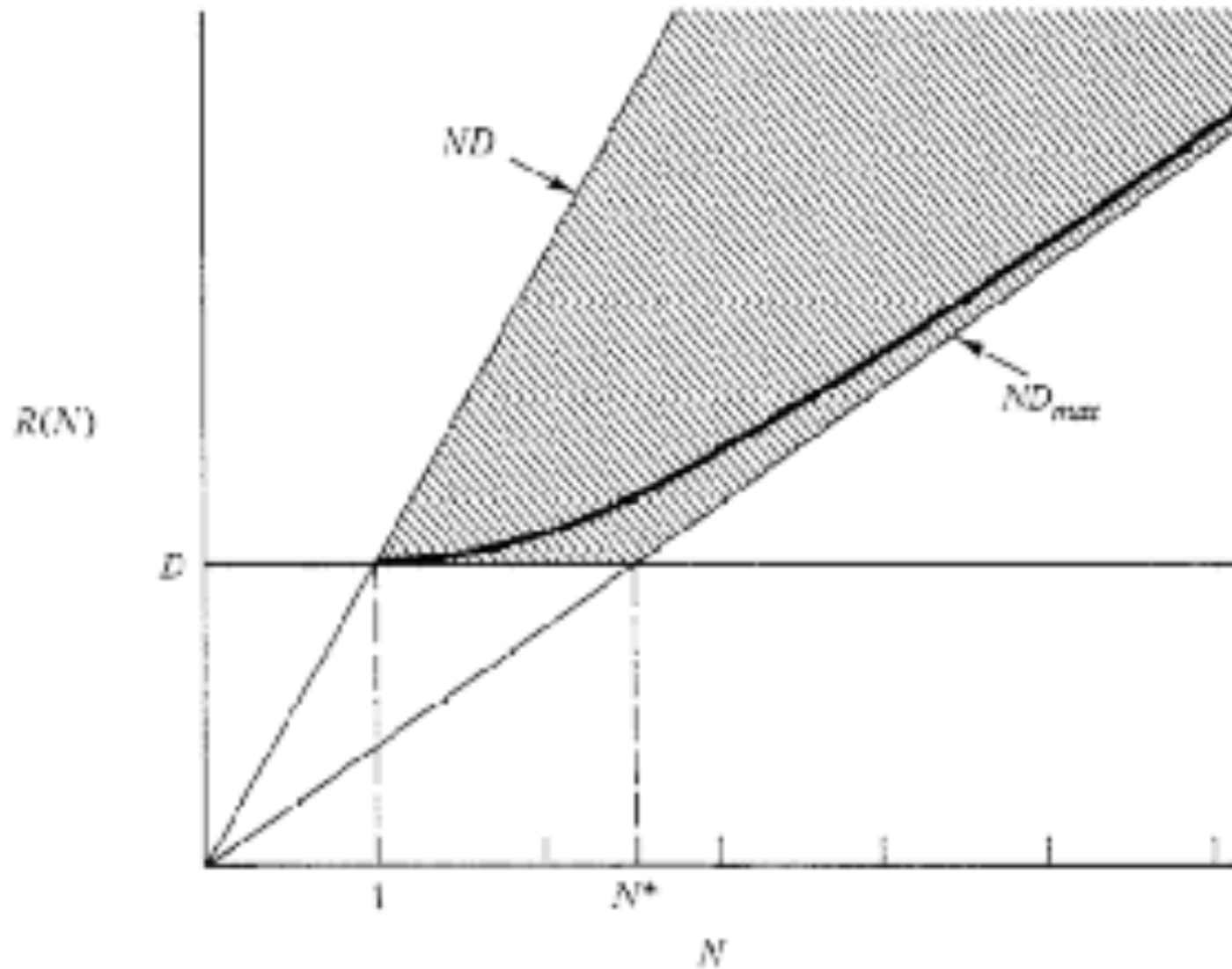
$$\max(D, ND_{\max}) \leq R(N) \leq ND$$

$N^* = D / D_{\max}$: ponto ótimo de operação

Limites Assintóticos para Cargas Batch (Fechado)



Limites Assintóticos para Cargas Batch (Fechado)



Aplicação dos Limites Assintóticos

Estudo de Caso

Uma companhia de seguros tem 20 sites distribuídos geograficamente onde a plataforma principal é X. Os tempos de resposta observados nos últimos tempos não têm sido aceitáveis, e a companhia decidiu realizar um upgrade. Os modelos Y30 e Y40 são capazes de executar as aplicações atuais e logo foram considerados. Após conversa com o vendedor, a companhia acredita que os Y30 iriam resultar em um ganho de "desempenho" de um fator de 1.5 a 2 (sobre os X) e que os Y40 levariam a uma melhora de desempenho de um fator de 2 a 3.5. Como o modelo Y30 é mais barato que o Y40, os gerentes estão inclinados a fazer o upgrade para o modelo Y30.

Aplicação dos Limites Assintóticos

Um estudo de modelagem foi iniciado. Através de medições nas plataformas existentes, resultados de avaliações anteriores e informações detalhadas dos fabricantes, descobriu-se que:

		service demands, seconds	
system		CPU	disk
X	(observed)	4.6	4.0
Y30	(estimated)	5.1	1.9
Y40	(estimated)	3.1	1.9

Desenvolva um modelo de limites assintóticos para avaliar as opções de upgrade

Aplicação dos Limites Assintóticos

- $K = 2$
- $D_{\max} = 4.6$ (X) $D_{\max} = 5.1$ (Y30) $D_{\max} = 3.1$ (Y40) (CPU)
- $D = 8.6$ (X) $D = 7.0$ (Y30) $D = 5.0$ (Y40)
- Tipo da classe de clientes: interativo/terminal
- $Z = 60$ segundos (medido)

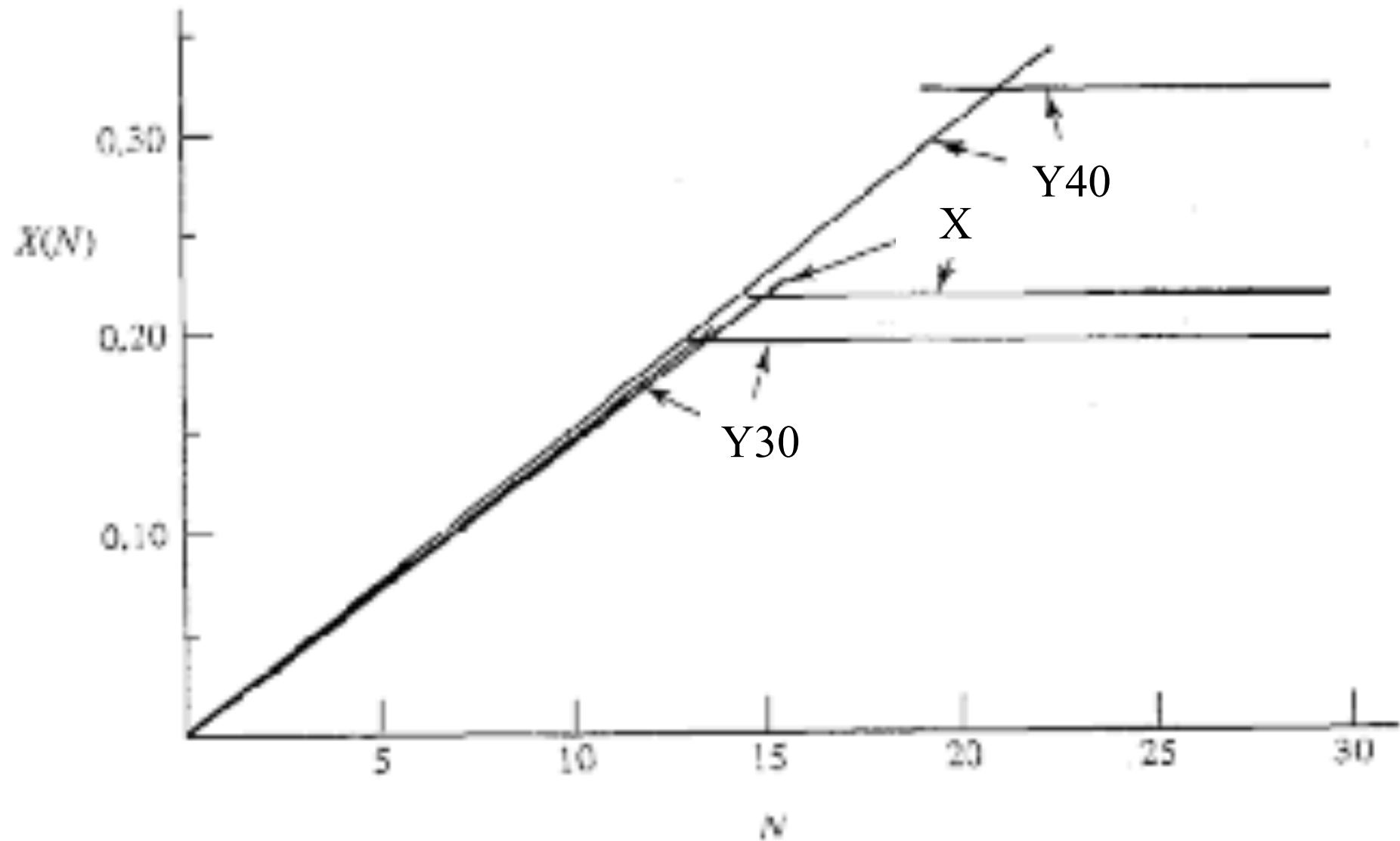
Aplicar limites para cada modelo (X, Y30 e Y40)

X

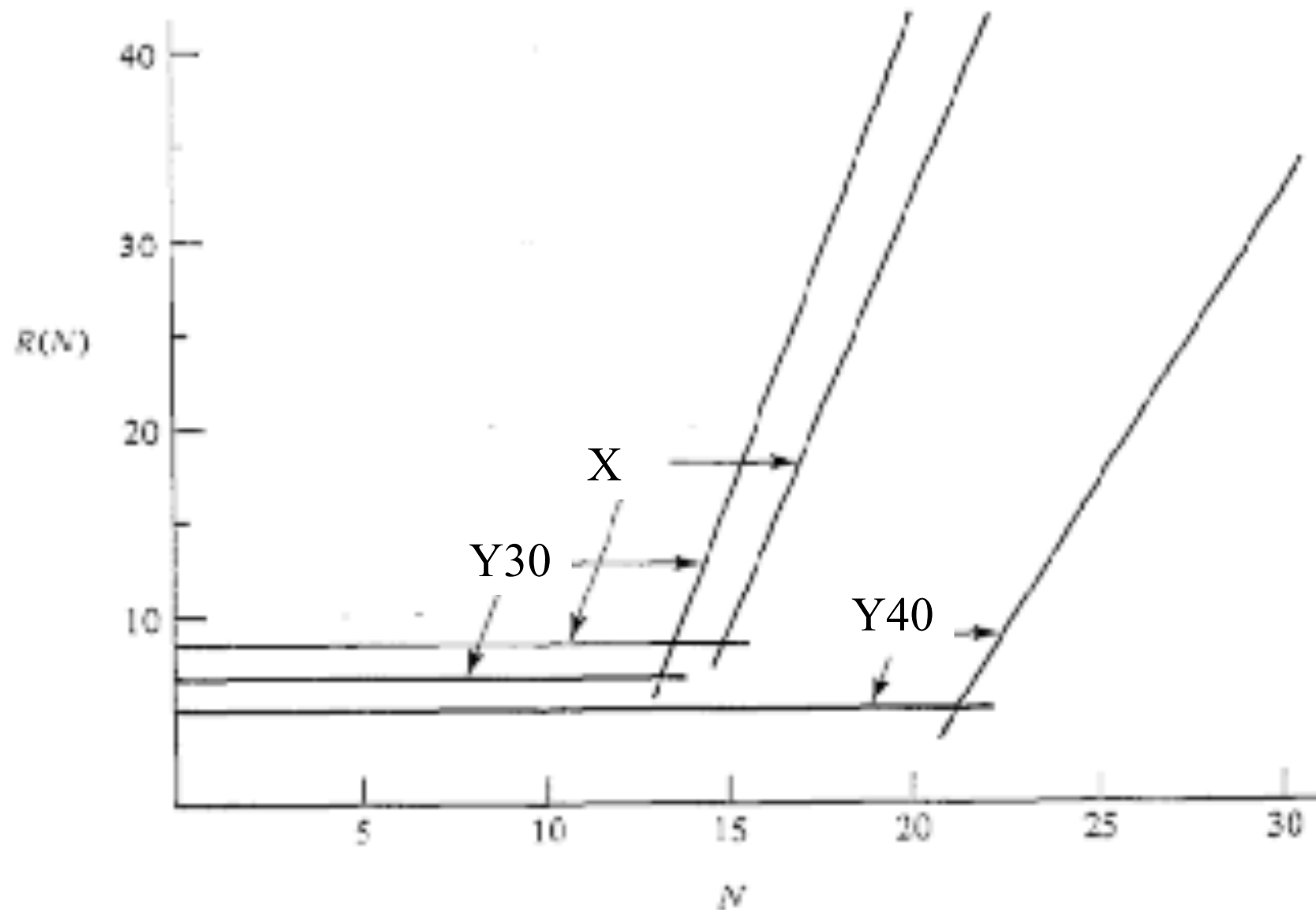
$$N / (ND + Z) \leq X(N) \leq \min(1/D_{\max}, N / (D+Z))$$
$$N / (8.6N+60) \leq X(N) \leq \min(1/4.6, N / 68.6)$$

$$\max(D, ND_{\max} - Z) \leq R(N) \leq ND$$
$$\max(8.6, 4.6N - 60) \leq R(N) \leq 8.6N$$

Aplicação dos Limites Assintóticos



Aplicação dos Limites Assintóticos



Aplicação dos Limites Assintóticos

- Medições foram feitas para re-avaliar a inclusão de Y30s no upgrade. Estas medições confirmaram que o desempenho dos Y30s seria pior que o dos Xs quando o número de terminais era aproximadamente 15 ou mais. O ganho para cargas leves é marginal. Logo não havia razão de incluir os Y30s no upgrade.
- Sem a modelagem dos limites, a companhia poderia ter seguido a opinião do vendedor e realizado o upgrade (desastroso) para Y30s

Aplicação dos Limites Assintóticos

- Com o upgrade para Y40 decidido, qual o aumento na carga máxima permitida?

$N^* = (D + Z) / D_{\max}$: ponto ótimo de operação

$$N^*_x = (8.6 + 60) / 4.6 = 14.91$$

$$N^*_{y40} = (5 + 60) / 3.1 = 20.96$$

$$(20.96 - 14.91) / 14.91 = 41\%$$

Aplicação dos Limites Assintóticos

Estudo de Caso 2

Seja um servidor de banco de dados com 1 CPU e 3 discos. Assuma que todas as transações do banco de dados têm demandas por recursos similares e o servidor está executando sob uma carga de transações estável. Medições realizadas durante uma hora mostraram que foram processadas 13680 transações e que:

	Número de IO por seg.	Utilização
Disco 1	32	0.3
Disco 2	36	0.41
Disco 3	50	0.54

Qual o tempo de serviço médio e a demanda média de cada transação em cada disco ? Qual o throughput do servidor ?

Aplicação dos Limites Assintóticos

Estudo de Caso 2

$$X_{\text{disco1}} = 32 \quad U_{\text{disco1}} = 0.30 \quad S_{\text{disco1}} = U_{\text{disco1}}/X_{\text{disco1}} = 0.0094$$

$$X_{\text{disco2}} = 36 \quad U_{\text{disco2}} = 0.41 \quad S_{\text{disco2}} = U_{\text{disco2}}/X_{\text{disco2}} = 0.0114$$

$$X_{\text{disco3}} = 50 \quad U_{\text{disco3}} = 0.54 \quad S_{\text{disco3}} = U_{\text{disco3}}/X_{\text{disco3}} = 0.0108$$

$$X = 13680/3600 = 3.8 \text{ transacoes/seg}$$

$$D_{\text{disco1}} = U_{\text{disco1}}/X = 0.0789 \quad D_{\text{disco2}} = U_{\text{disco2}}/X = 0.1079$$

$$D_{\text{disco3}} = U_{\text{disco3}}/X = 0.1421$$

Aplicação dos Limites Assintóticos

Estudo de Caso 2

Sabendo que a demanda média de uma transação pela CPU é 0.092, responda:

- 1) Qual o gargalo do sistema?
- 2) Quais deveriam ser as demandas nos três discos para se obter o maior throughput possível, mantendo constante a demanda total sobre todos os 3 discos (mantendo mesma carga total) ?

Aplicação dos Limites Assintóticos

Estudo de Caso 2

Sabendo que a demanda média de uma transação pela CPU é 0.092, responda:

1) Qual o gargalo do sistema?

$$D_{\text{disco1}} = 0.0789, \quad D_{\text{disco2}} = 0.1079, \quad D_{\text{disco3}} = 0.1421 \\ D_{\text{cpu}} = 0.092$$

O disco 3 é o gargalo do sistema e candidato ao próximo upgrade

Aplicação dos Limites Assintóticos

Estudo de Caso 2

Sabendo que a demanda média de uma transação pela CPU é 0.092, responda:

- 2) Quais deveriam ser as demandas nos três discos para se obter o maior throughput possível, mantendo constante a demanda total sobre todos os 3 discos (mantendo mesma carga total) ?

$$D_{\text{disco}} = D_{\text{disco1}} + D_{\text{disco2}} + D_{\text{disco3}} = 0.0789 + 0.1079 + 0.1421 = 0.3289$$
$$D_{\text{disco}}/3 = 0.1097 \text{ seg.}$$

$$X \leq 1/D_{\text{max}} = 1/D_{\text{disco}} = 9.12 \text{ transacoes/seg}$$

$$\text{Antes } X \leq 1/D_{\text{disco3}} = 7.04 \text{ transacoes/seg}$$

Mudanca leva a aumento de 29.5%

Aplicação dos Limites Assintóticos

Efeito da Remoção de um Bottleneck

Ex: sistema com 3 centros de serviços e com uma carga interativa com think time médio igual a 15 segundos. As demandas nos três centros são 5, 4 e 3 segundos.

O que acontece se a carga no dispositivo bottleneck é reduzida através da substituição por um dispositivo mais rápido ou deslocando parte da carga para outro dispositivo?

Qual o impacto de melhorar o desempenho de um dispositivo que não é o bottleneck primário (bottleneck secundário)?

Aplicação dos Limites Assintóticos

Efeito da Remoção de um Bottleneck

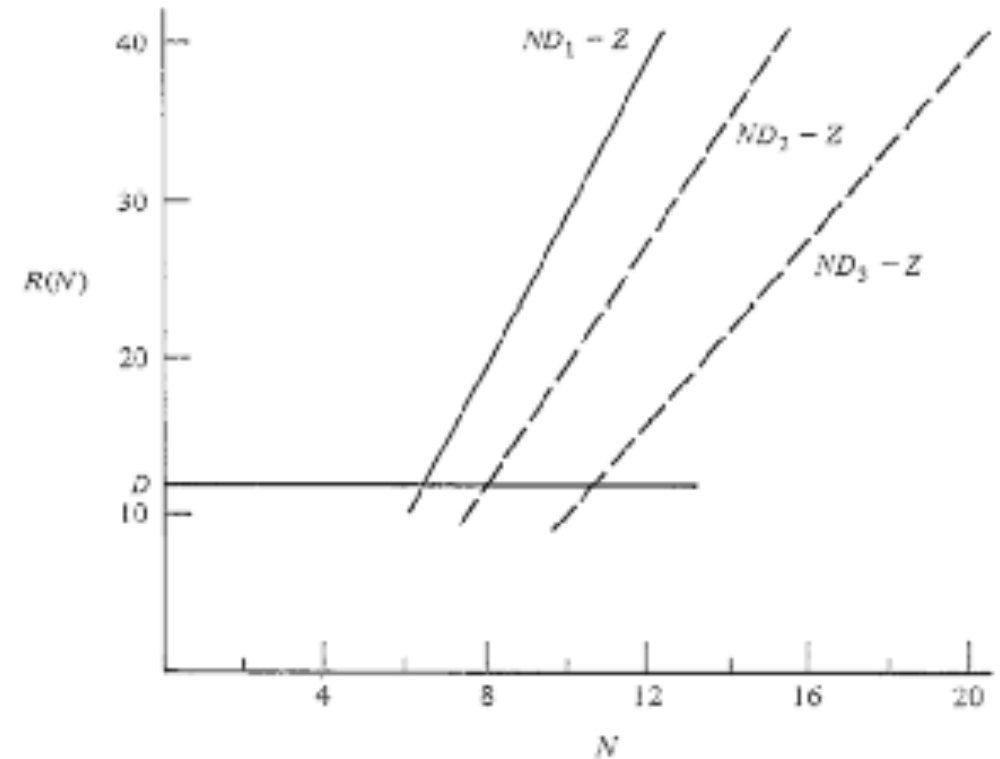
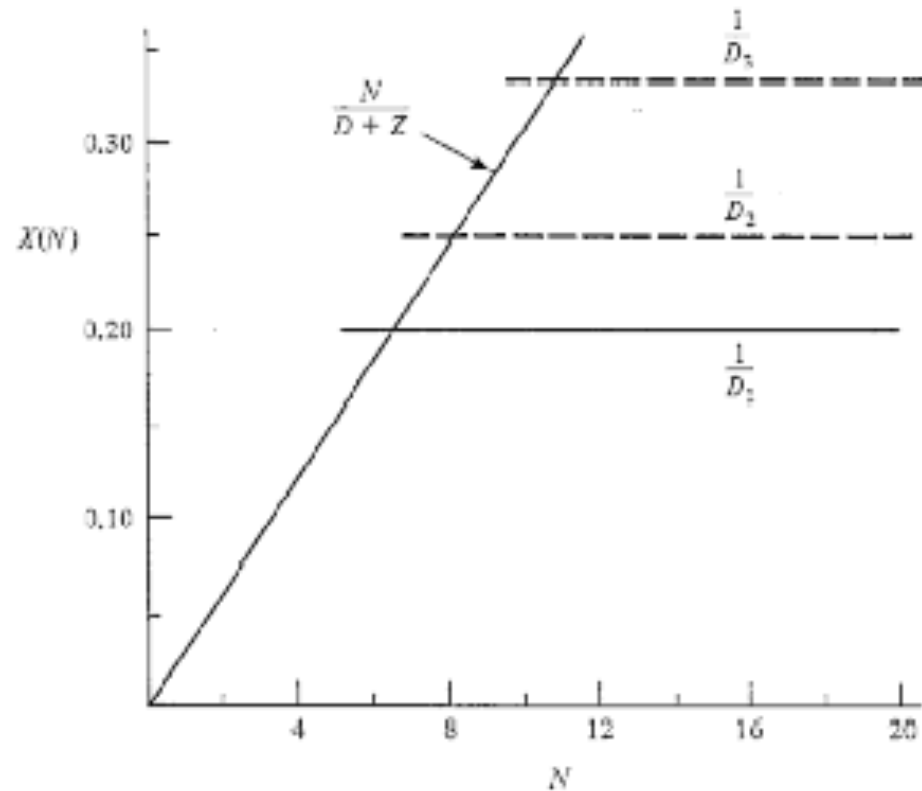
Ex: sistema com 3 centros de serviços e com uma carga interativa com think time médio igual a 15 segundos. As demandas nos três centros são 5, 4 e 3 segundos.

$$D_1 = 5, \quad D_2 = 4, \quad D_3 = 3, \quad D = 12, \quad Z = 15$$

O que acontece se a carga no dispositivo bottleneck é reduzida através da substituição por um dispositivo mais rápido ou deslocando parte da carga para outro dispositivo?

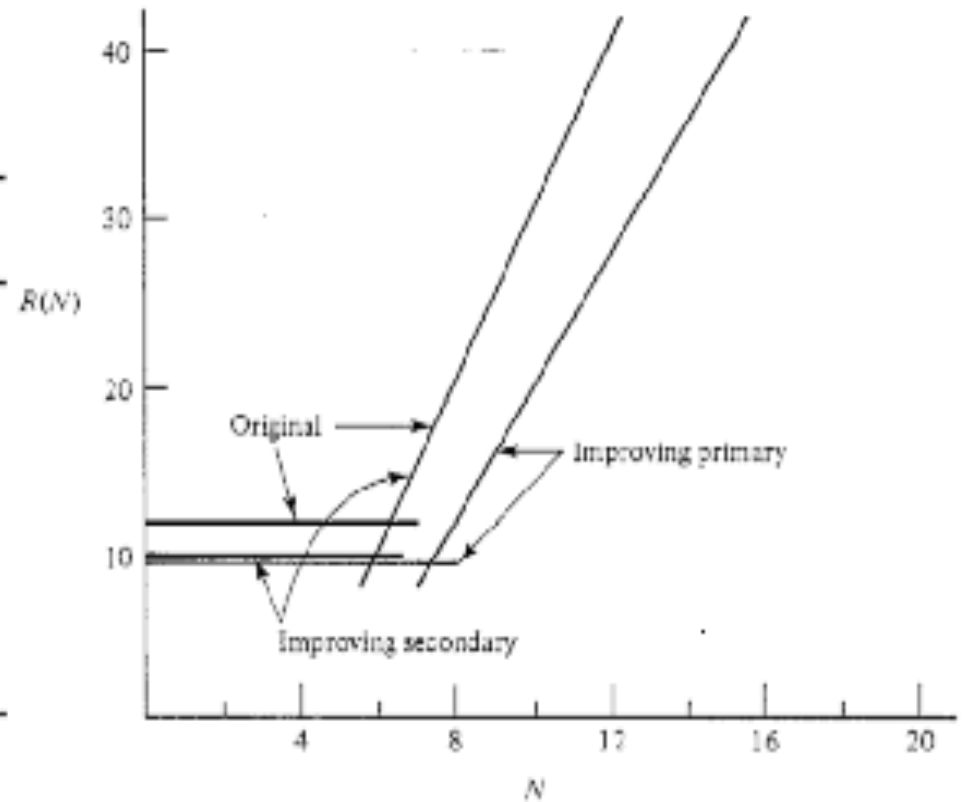
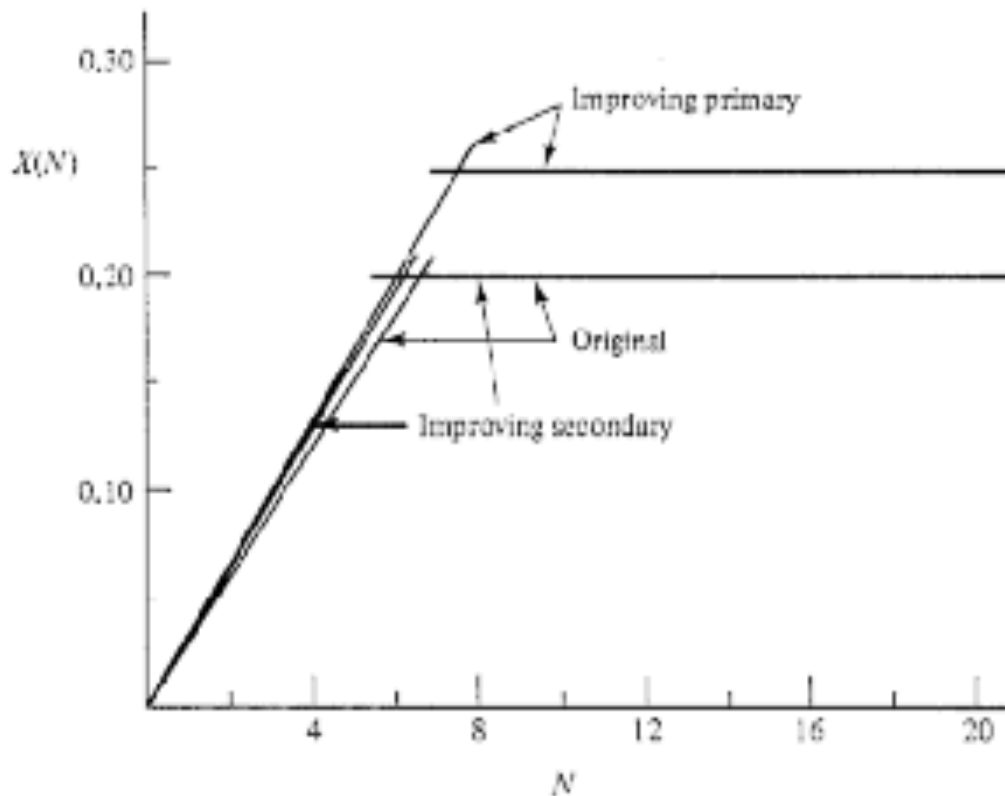
Qual o impacto de melhorar o desempenho de um dispositivo que não é o bottleneck primário (bottleneck secundário)?

Aplicação dos Limites Assintóticos



- Melhorar desempenho de bottleneck secundários:
Impacto mínimo no caso de carga leve e nenhum para pesada
- Ao remover o bottleneck principal, o dispositivo que tem a segunda maior demanda (originalmente) se torna o bottleneck

Aplicação dos Limites Assintóticos



- Melhorar desempenho de bottleneck secundários:
Impacto mínimo no caso de carga leve e nenhum para pesada
- Ao remover o bottleneck principal, o dispositivo que tem a atual maior demanda se torna o bottleneck

Aplicação dos Limites Assintóticos

Impacto de Modificações em um Sistema Existente

Considere um sistema interativo no qual as seguintes medições foram feitas:

$T = 900$ segundos

$Z = 15$ segundos

$B_1 = 400$ segundos (CPU)

$C = 200$ interações

$B_2 = 100$ segundos (disco lento)

$C_2 = 2000$

$B_3 = 600$ segundos (disco rápido)

$C_3 = 20000$

Avalie os ganhos de desempenho com os seguintes upgrades:

- 1) Substituir a CPU por outra duas vezes mais rápida
- 2) Deslocar alguns arquivos do disco mais rápido para o disco mais lento, balanceando a carga
- 3) Adicionar um segundo disco rápido para servir metade da carga do disco mais ocupado atualmente
- 4) As três mudanças feitas juntas

Aplicação dos Limites Assintóticos

Calcular métricas de carga

Demandas: $D_i = B_i/C$

$$D_1 = 400/200 = 2.0, D_2 = 0.5, D_3 = 3.0$$

Número de Visitas: $V_i = C_i/C$

$$V_2 = 2000/200 = 10, V_3 = 100$$

Tempo de Serviço: $S_i = B_i/C_i$

$$S_2 = 100/2000 = 0.05, S_3 = 0.03$$

Aplicação dos Limites Assintóticos

Opção 1: Substituir a CPU por outra duas vezes mais rápida

Dividir demanda de CPU por 2

$$D'_1 = 2.0 / 2 = 1.0$$

Opção 2: Balancear demanda entre os dois discos

Considerar somente efeito primário que é a velocidade do disco e ignorar efeitos secundários com o tamanho de bloco

$$V_2 + V_3 = 110 = V'_2 + V'_3$$

Já que $S_2 = 0.05$, $S_3 = 0.03$:

$$V_2 S_2 / 0.05 + V_3 S_3 / 0.03 = 110 \quad D'_2 = V'_2 S_2 = V'_3 S_3 = D'_3$$

$$D'_2 [1/0.05 + 1/0.03] = 110 \quad D'_2 = D'_3 = 2.06$$

$$V'_2 = 41 \text{ e } V'_3 = 69$$

Aplicação dos Limites Assintóticos

Opção 3: Adicionar 3o disco rápido para servir metade da carga do disco mais ocupado

Dividir demanda do disco mais rápido por 2

$$K = 4$$

$$D'_3 = D'_4 = 3.0 / 2 = 1.5$$

Opção 4: 1 + 2 + 3

$$D'_1 = 1.0$$

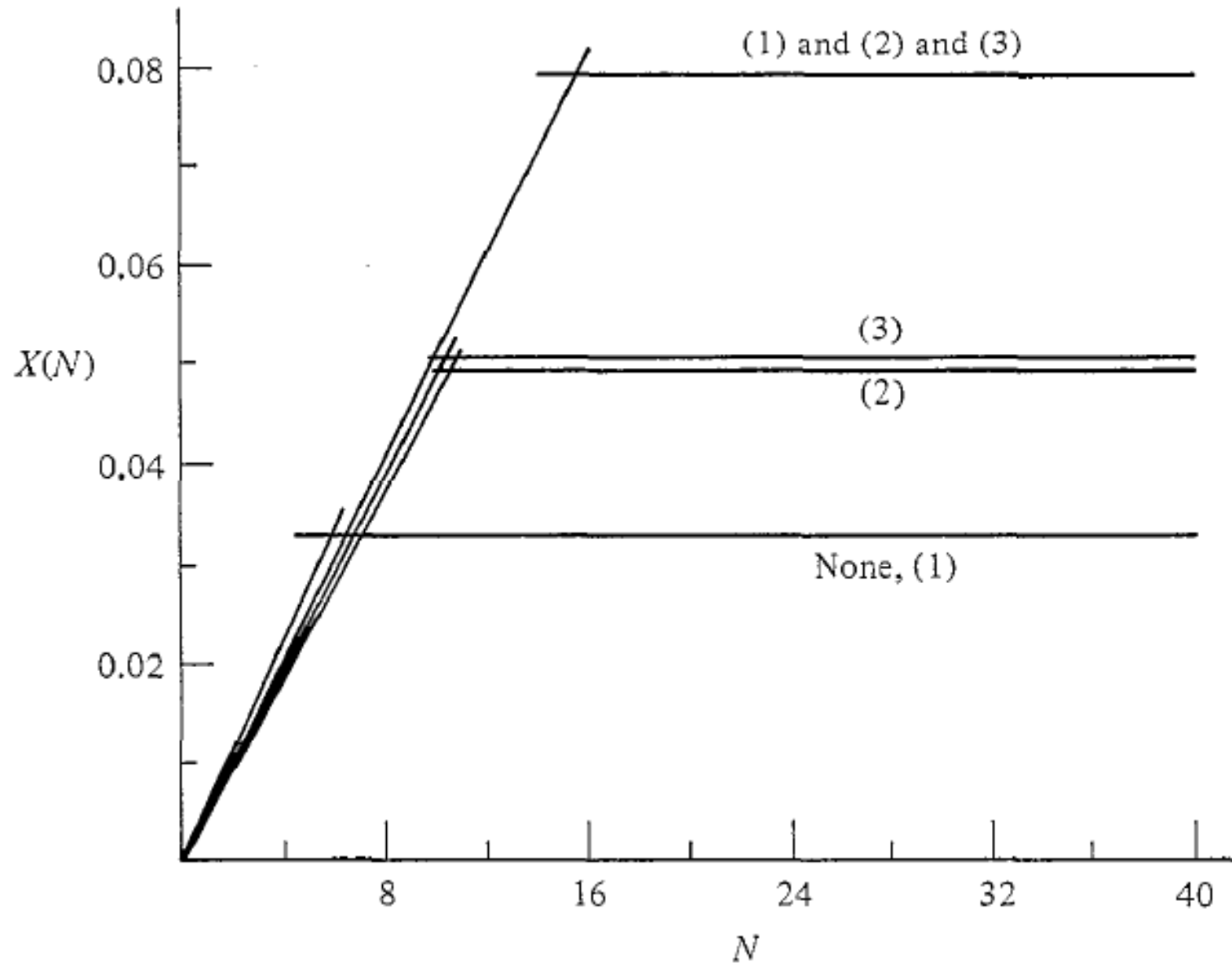
$$V_2 + V_3 + V_4 = 110$$

$$\text{Já que } S_2 = 0.05, S_3 = S_4 = 0.03$$

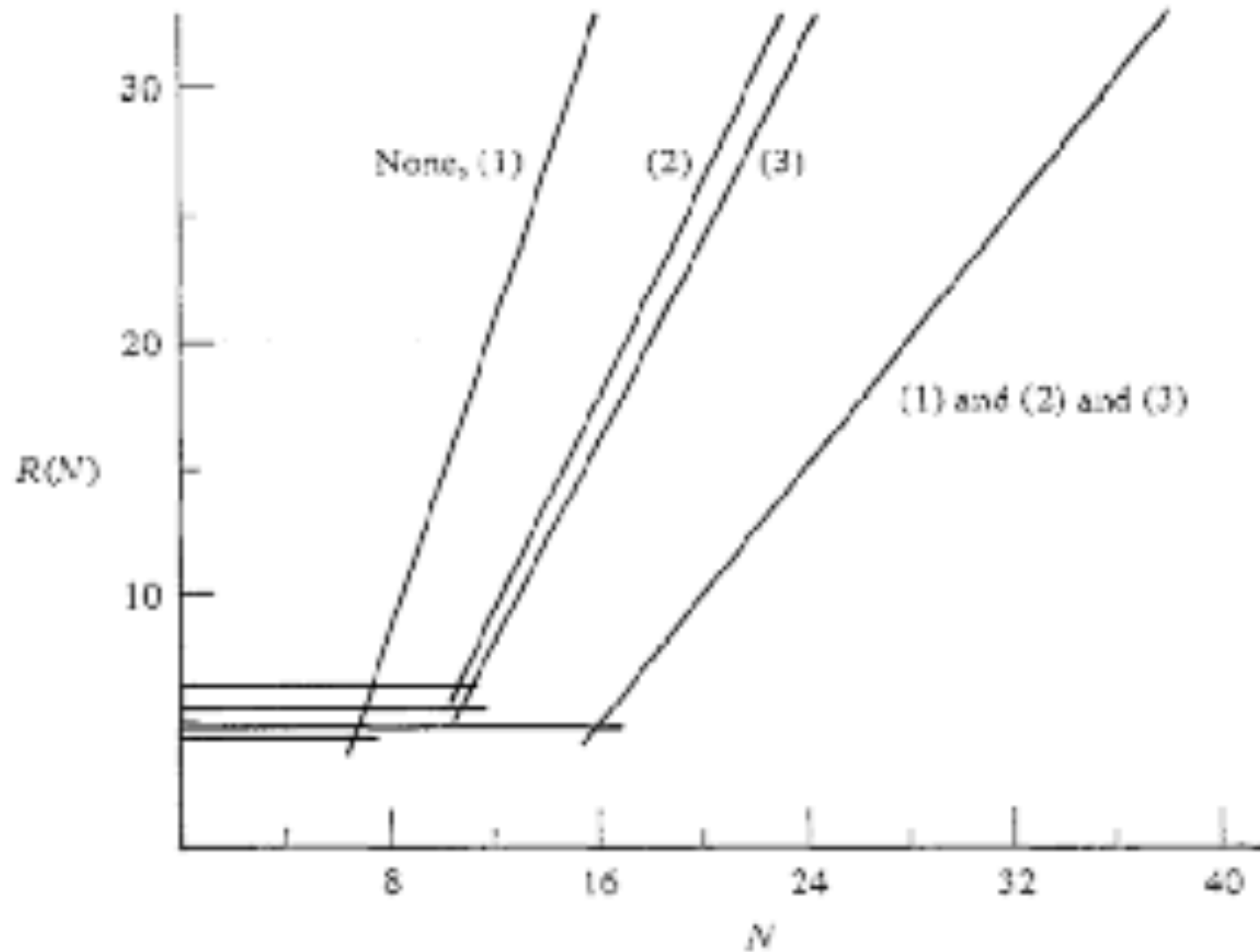
$$V'_2 S_2 / 0.05 + V'_3 S_3 / 0.03 + V'_4 S_4 / 0.03 = 110 \quad D'_2 = D'_3 = D'_4$$

$$D'_2 [1/0.05 + 1/0.03 + 1/0.03] = 110 \quad D'_2 = D'_3 = D'_4 = 1.27$$

Aplicação dos Limites Assintóticos



Aplicação dos Limites Assintóticos



Limites Sistemas Balanceados

- Mais apertados mas exigem mais cálculo
- Baseados em sistemas que estão *balanceados*:
 - $D_1 = D_2 = D_3 = \dots = D_k$
- Premissa: redes separáveis
 - Cada centro pode ser analisado independentemente
 - Várias características (cap. 6 do Lazowska)
- Utilização de cada centro k em sistemas balanceados dada por:
 - $U_k(N) = N / (N + K - 1)$

Limites Sistemas Balanceados: Aplicação

- Seja um modelo de um sistema onde $D_{\max}$, D_{ave} e $D_{\min}$ representem as demandas máxima, média e mínima por serviço nos centros, obtenha limites para throughput X e tempo de resposta R
- Lembrar que:

$$U_k = X(N)D_k$$

$$X(N) = \frac{U_k}{D_k} = \frac{N}{N + K - 1} \times \frac{1}{D_k}$$

Limites Sistemas Balanceados: Carga Batch v.1

- Limites inferior e superior para throughput do sistema dado pelos throughputs dos sistemas balanceados com demanda $D_{\min}$ e $D_{\max}$.

$$\frac{N}{N + K - 1} \times \frac{1}{D_{\max}} \leq X(N) \leq \frac{N}{N + K - 1} \times \frac{1}{D_{\min}}$$

- Aplicando Lei de Little, temos:

$$(N + K - 1)D_{\min} \leq R(N) \leq (N + K - 1)D_{\max}$$

Limites Sistemas Balanceados: Carga Batch v.2

- Limites mais apertados
- De todos os sistemas com demanda total D :
 - maior throughput: sistema no qual todos centros têm demanda $D_{ave} = D/K$

$$X(N) \leq \frac{N}{N + K - 1} \times \frac{1}{D_{ave}} = \frac{N}{D + (N - 1)D_{ave}}$$

- menor throughput: sistema que tem D/D_{max} centros com demanda D_{max} , cada, e $D_k = 0$ nos outros centros.

$$\frac{N}{N + \frac{D}{D_{max}} - 1} \times \frac{1}{D_{max}} = \frac{N}{D + (N - 1)D_{max}} \leq X(N)$$

Limites Sistemas Balanceados: Carga Batch v.2

- Limites para throughput:

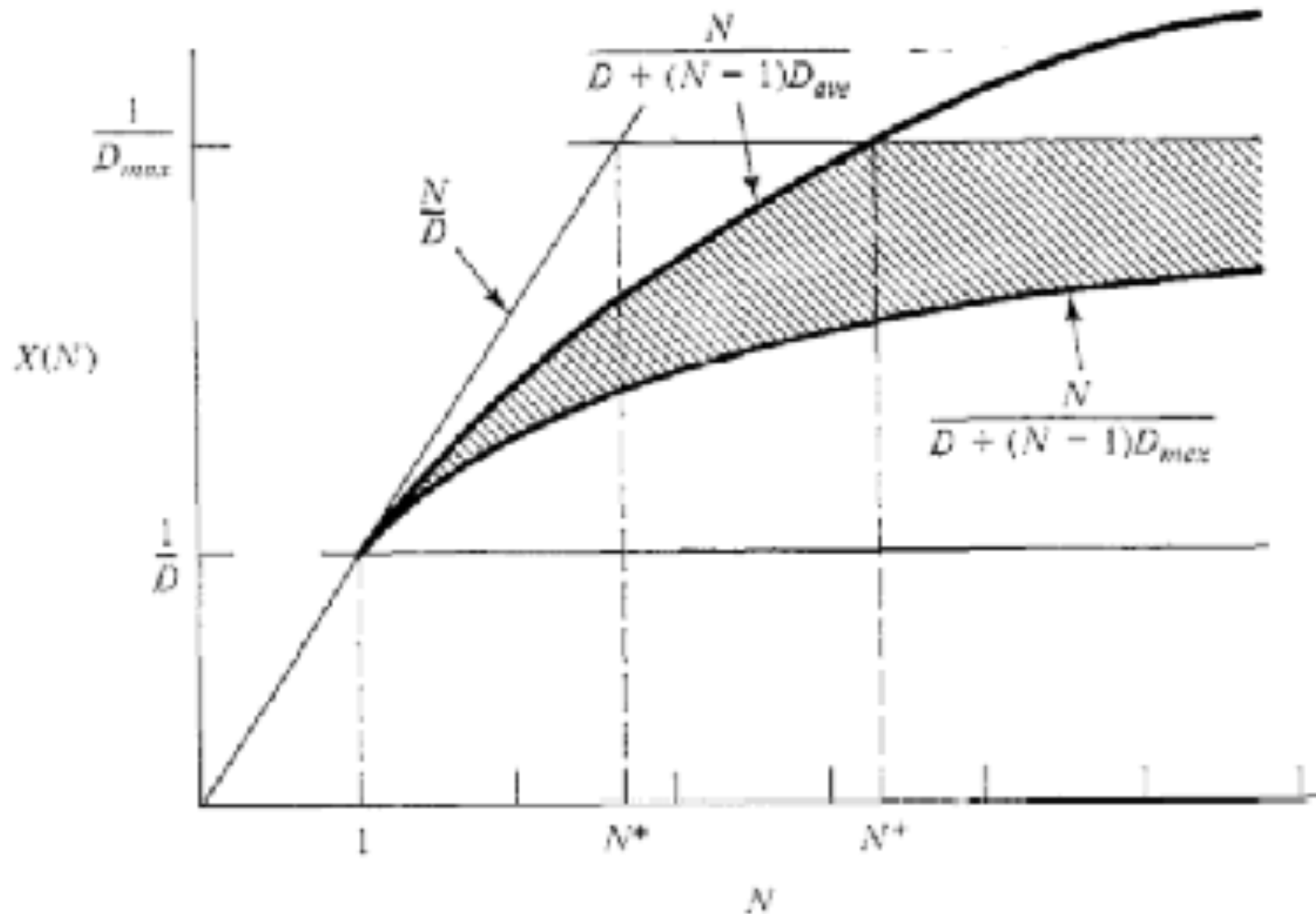
$$\frac{N}{D+(N-1)D_{ave}} \leq X(N) \leq \frac{N}{D+(N-1)D_{ave}}$$

- Aplicando Little

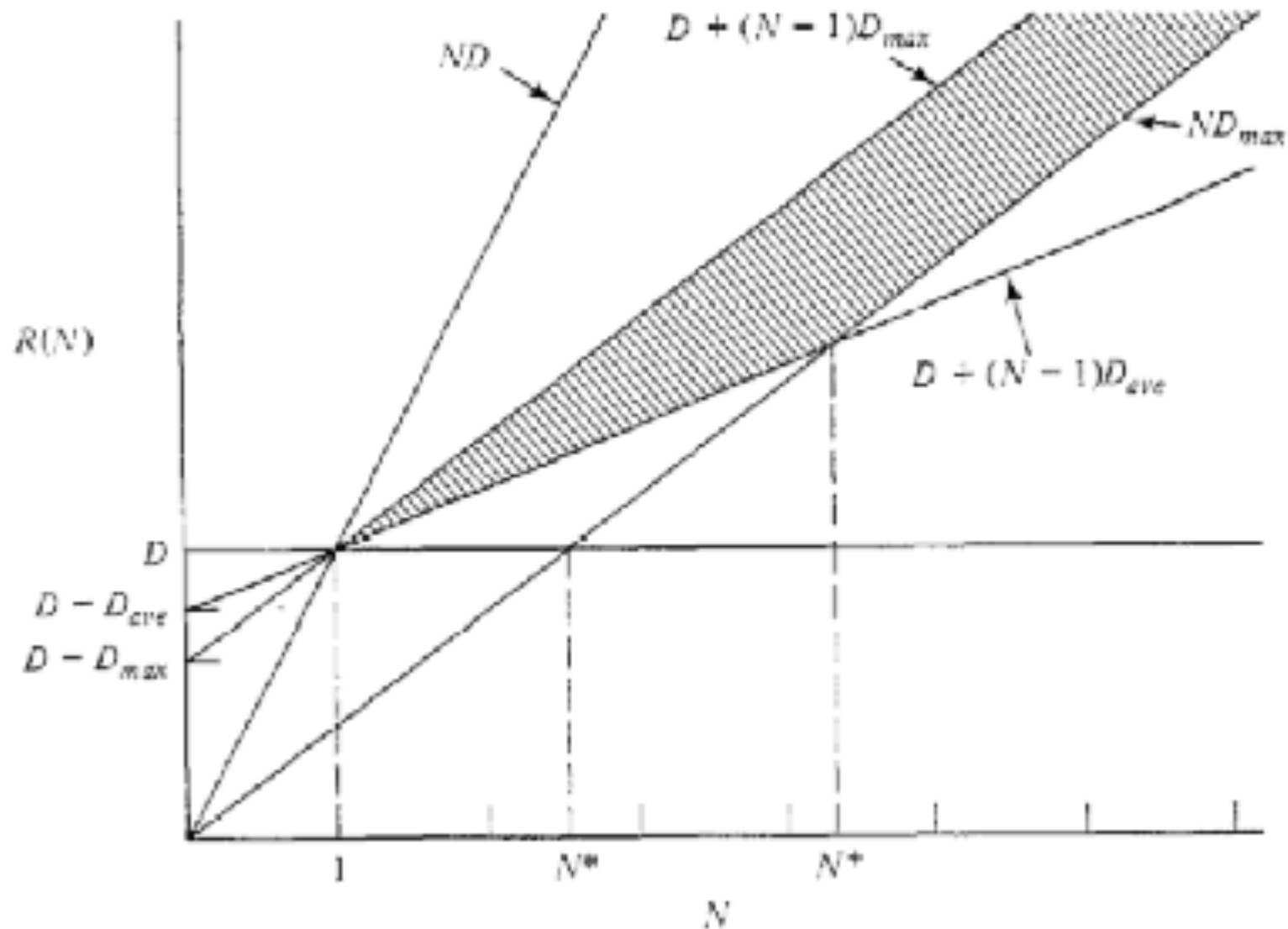
$$D + (N - 1)D_{ave} \leq R(N) \leq D + (N - 1)D_{max}$$

- N^+ : ponto onde o limite otimista para sistema balanceado intersecta o limite otimista assintótico
 - $N > N^+$: limite para sistema balanceado = limite assintótico

Limites Sistemas Balanceados: Carga Batch



Limites Sistemas Balanceados: Carga Batch



Límites Sistemas Balanceados: Sumário

	workload type	bounds
X	batch	$\frac{N}{D + (N-1)D_{\max}} \leq X(N)$ $\leq \min \left(\frac{1}{D_{\max}}, \frac{N}{D + (N-1)D_{\text{ave}}} \right)$
	terminal	$\frac{N}{D + Z + \frac{(N-1)D_{\max}}{1 + Z/(ND)}} \leq X(N)$ $\leq \min \left(\frac{1}{D_{\max}}, \frac{N}{D + Z + \frac{(N-1)D_{\text{ave}}}{1 + Z/D}} \right)$
	transaction	$X(\lambda) \leq 1 / D_{\max}$
R	batch	$\max (ND_{\max}, D + (N-1)D_{\text{ave}}) \leq R(N)$ $\leq D + (N-1)D_{\max}$
	terminal	$\max (ND_{\max} - Z, D + \frac{(N-1)D_{\text{ave}}}{1 + Z/D}) \leq R(N)$ $\leq D + \frac{(N-1)D_{\max}}{1 + Z/(ND)}$
	transaction	$\frac{D}{1 - \lambda D_{\text{ave}}} \leq R(\lambda) \leq \frac{D}{1 - \lambda D_{\max}}$