Dynamic Content Allocation for Cloud-assisted Service of Periodic Workloads

György Dán

Royal Institute of Technology (KTH)

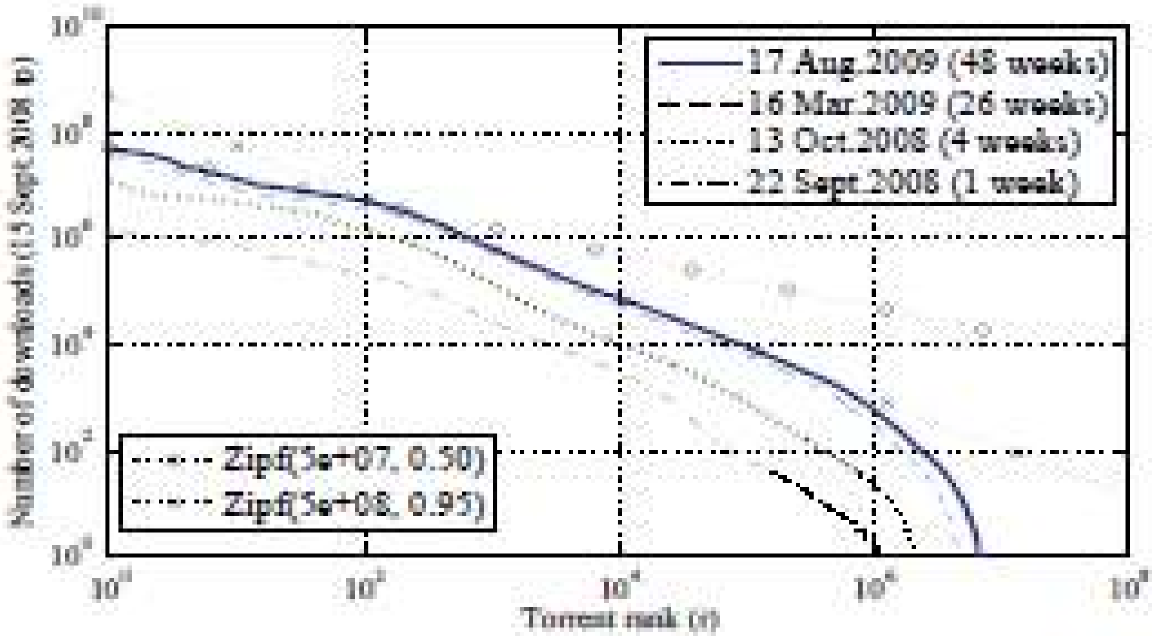
Niklas Carlsson

Linköping University

LiU EXPANDING REALITY

***@ IEEE INFOCOM 2014*, Toronto, Canada, April/May 2014**

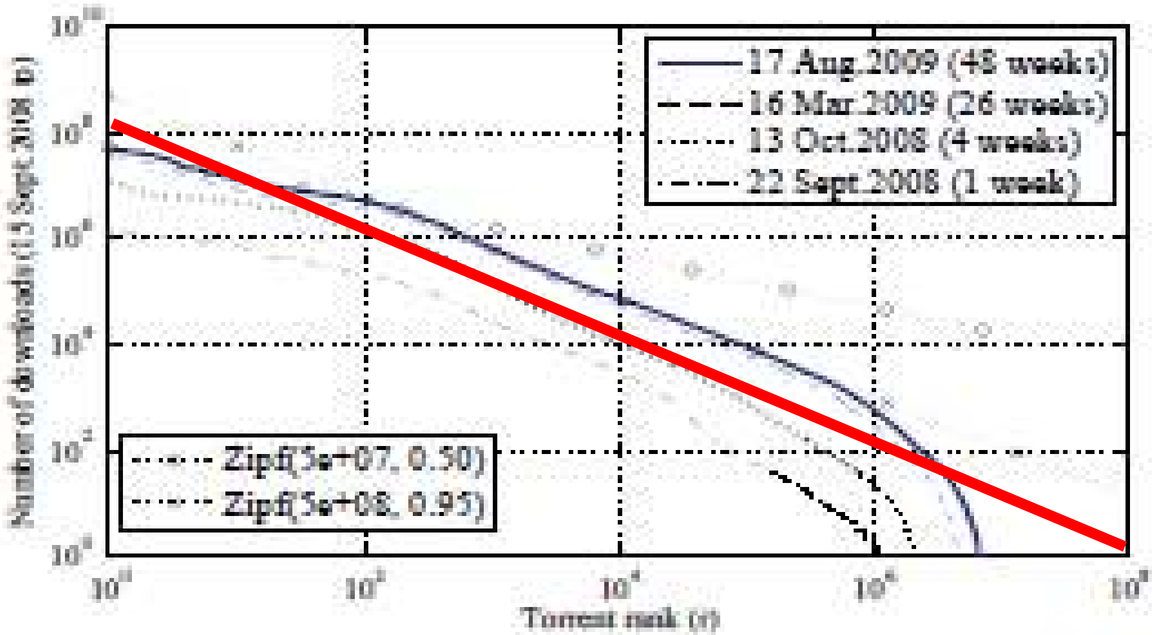
Internet Content Delivery



From: *Dan and Carlsson, "Power-laws Revisited: A Large Scale Measurement Study of Peer-to-Peer Content Popularity", Proc. IPTPS 2010.*

- Large amounts of data with varying popularity
- Multi-billion market ($8B to $20B, 2012-2015)
  - Goal: Minimize content delivery costs
- Migration to cloud data centers
-

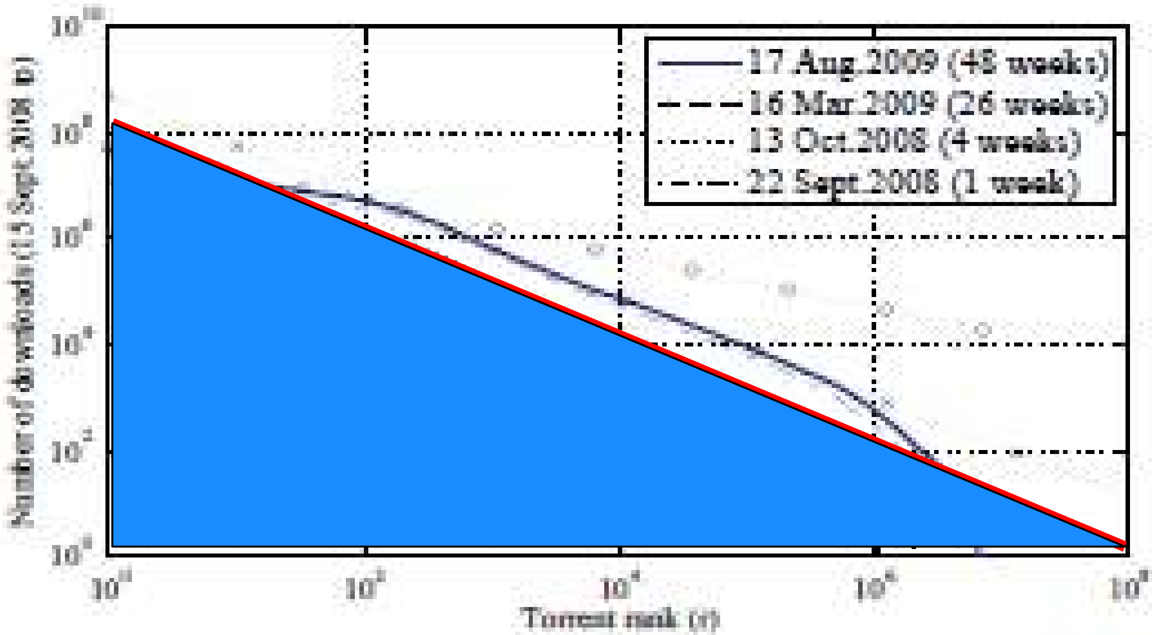
Internet Content Delivery



From: *Dan and Carlsson, "Power-laws Revisited: A Large Scale Measurement Study of Peer-to-Peer Content Popularity", Proc. IPTPS 2010.*

- Large amounts of data with varying popularity
- Multi-billion market ($8B to $20B, 2012-2015)
    - Goal: Minimize content delivery costs
- Migration to cloud data centers

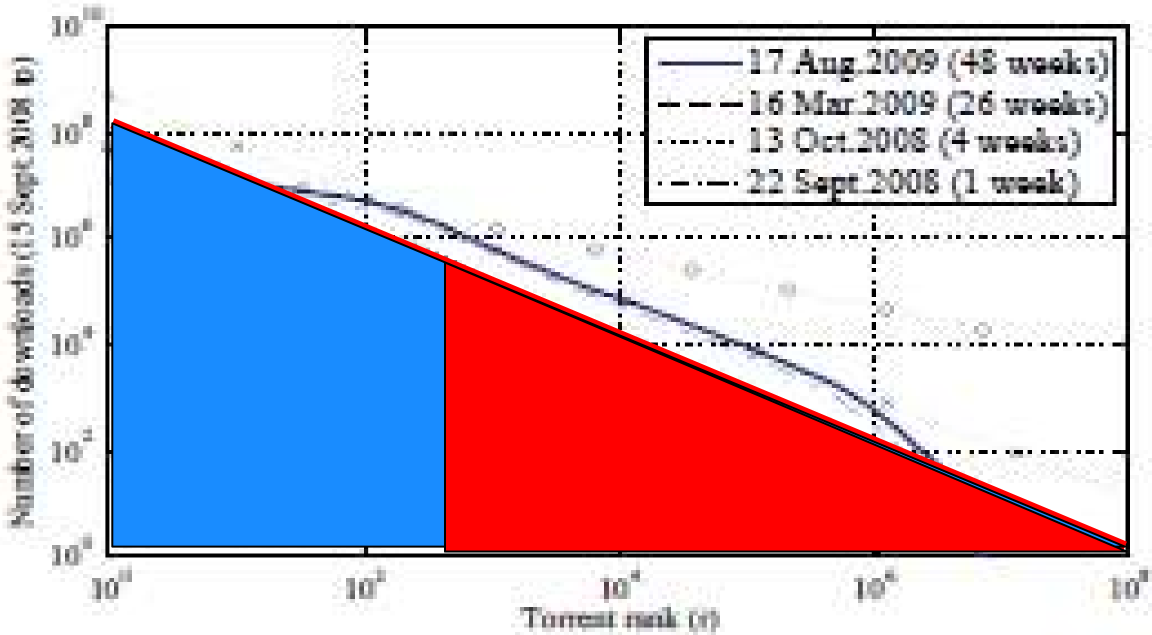
Internet Content Delivery



From: *Dan and Carlsson, "Power-laws Revisited: A Large Scale Measurement Study of Peer-to-Peer Content Popularity", Proc. IPTPS 2010.*

- Large amounts of data with varying popularity
- Multi-billion market ($8B to $20B, 2012-2015)
  - Goal: Minimize content delivery costs
- Migration to cloud data centers

Internet Content Delivery

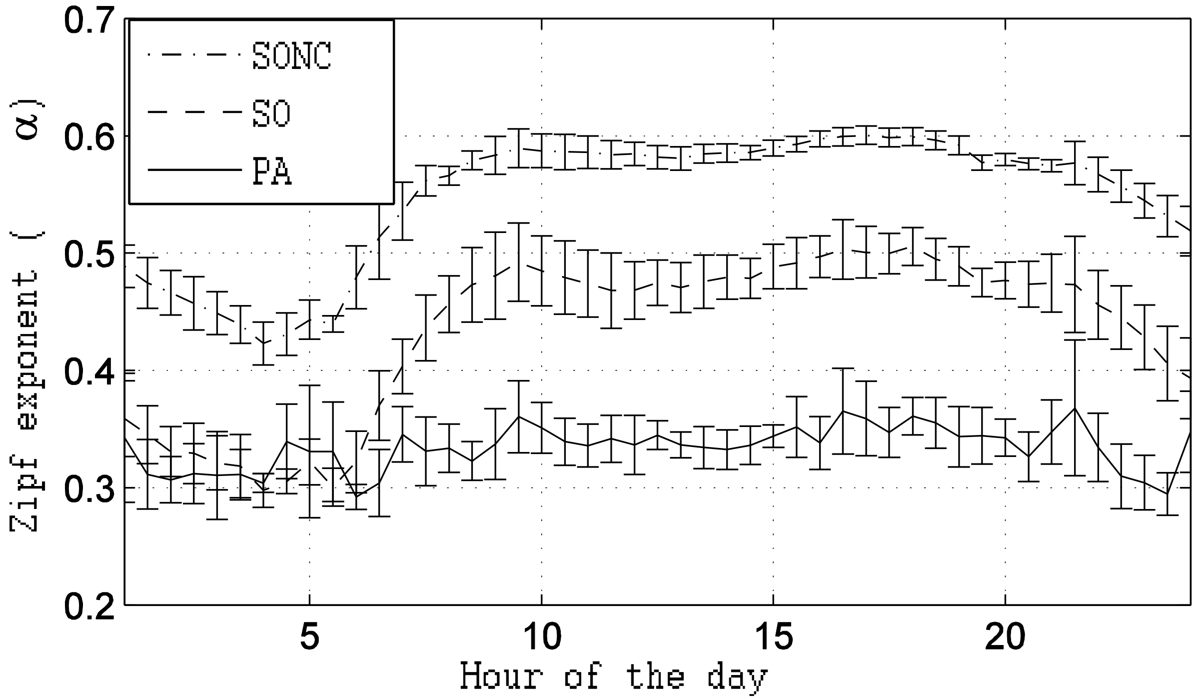


From: *Dan and Carlsson, "Power-laws Revisited: A Large Scale Measurement Study of Peer-to-Peer Content Popularity", Proc. IPTPS 2010.*

- Large amounts of data with varying popularity
- Multi-billion market ($8B to $20B, 2012-2015)
  - Goal: Minimize content delivery costs
- Migration to cloud data centers
-

Periodic Workloads

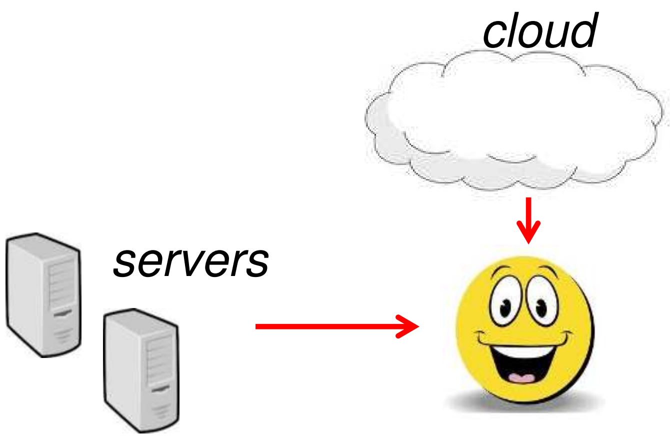
- Characterization of Spotify traces
- In addition to diurnal traffic volumes …
- … we found that also the Zipf exponent vary with time-of-day



Content Delivery

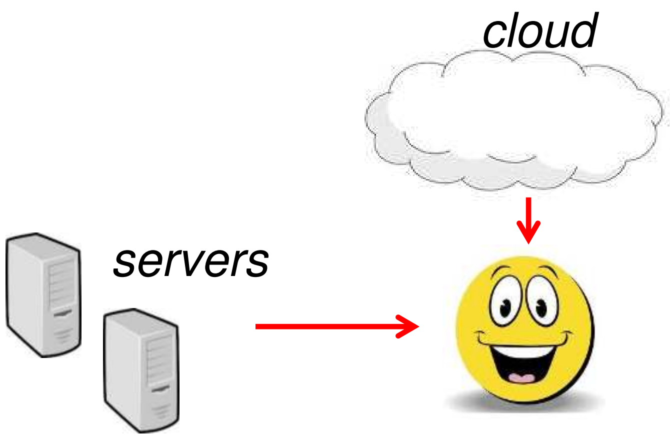
- Cloud-based delivery

- Dedicated infrastructure



Content Delivery

- Cloud-based delivery
  - Flexible computation, storage, and bandwidth
  - Pay per volume and access
- Dedicated infrastructure
  - Limited storage
  - Capped unmetered bandwidth
  - Potentially closer to the user



Content Delivery

- Cloud-based delivery
  - Flexible computation, storage, and bandwidth
  - Pay per volume and access
- Dedicated infrastructure
  - Limited storage
  - Capped unmetered bandwidth
  - Potentially closer to the user

Content Delivery

- Cloud-based delivery
  - Flexible computation, storage, and bandwidth
  - Pay per volume and access
- Dedicated infrastructure
  - Limited storage
  - Capped unmetered bandwidth
  - Potentially closer to the user

Content Delivery

- Cloud-based delivery
  - Flexible computation, storage, and bandwidth
  - Pay per volume and access
- Dedicated infrastructure
  - Limited storage
  - Capped unmetered bandwidth
  - Potentially closer to the user

Content Delivery

- Cloud-based delivery
  - Flexible computation, storage, and bandwidth
  - Pay per volume and access
- Dedicated infrastructure
  - Limited storage
  - Capped unmetered bandwidth
  - Potentially closer to the user

Cloud bandwidth elastic; however, flexible comes at premium …

High-level problem

- Minimize content delivery costs

	Bandwidth	Cost
Cloud-based	Elastic/flexible	$$$
Dedicated servers	Capped	$

-

High-level problem

- Minimize content delivery costs

	Bandwidth	Cost
Cloud-based	Elastic/flexible	$$$
Dedicated servers	Capped	$

How to get the best of two worlds?

High-level problem

- Minimize content delivery costs

	Bandwidth	Cost
Cloud-based	Elastic/flexible	$$$
Dedicated servers	Capped	$

- How to get the best out of two worlds?

-

High-level problem

- Minimize content delivery costs

	Bandwidth	Cost
Cloud-based	Elastic/flexible	$$$
Dedicated servers	Capped	$

- How to get the best out of two worlds?
  - Improved workload models and prediction enables prefetching …
-

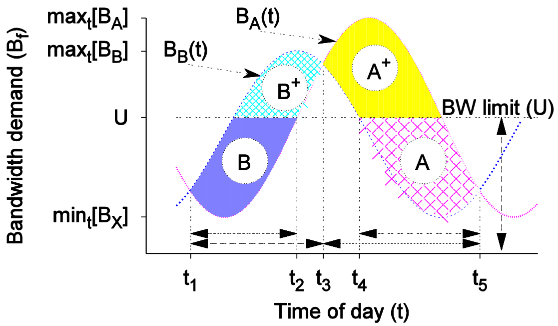
High-level problem

- Minimize content delivery costs

	Bandwidth	Cost
Cloud-based	Elastic/flexible	$$$
Dedicated servers	Capped	$

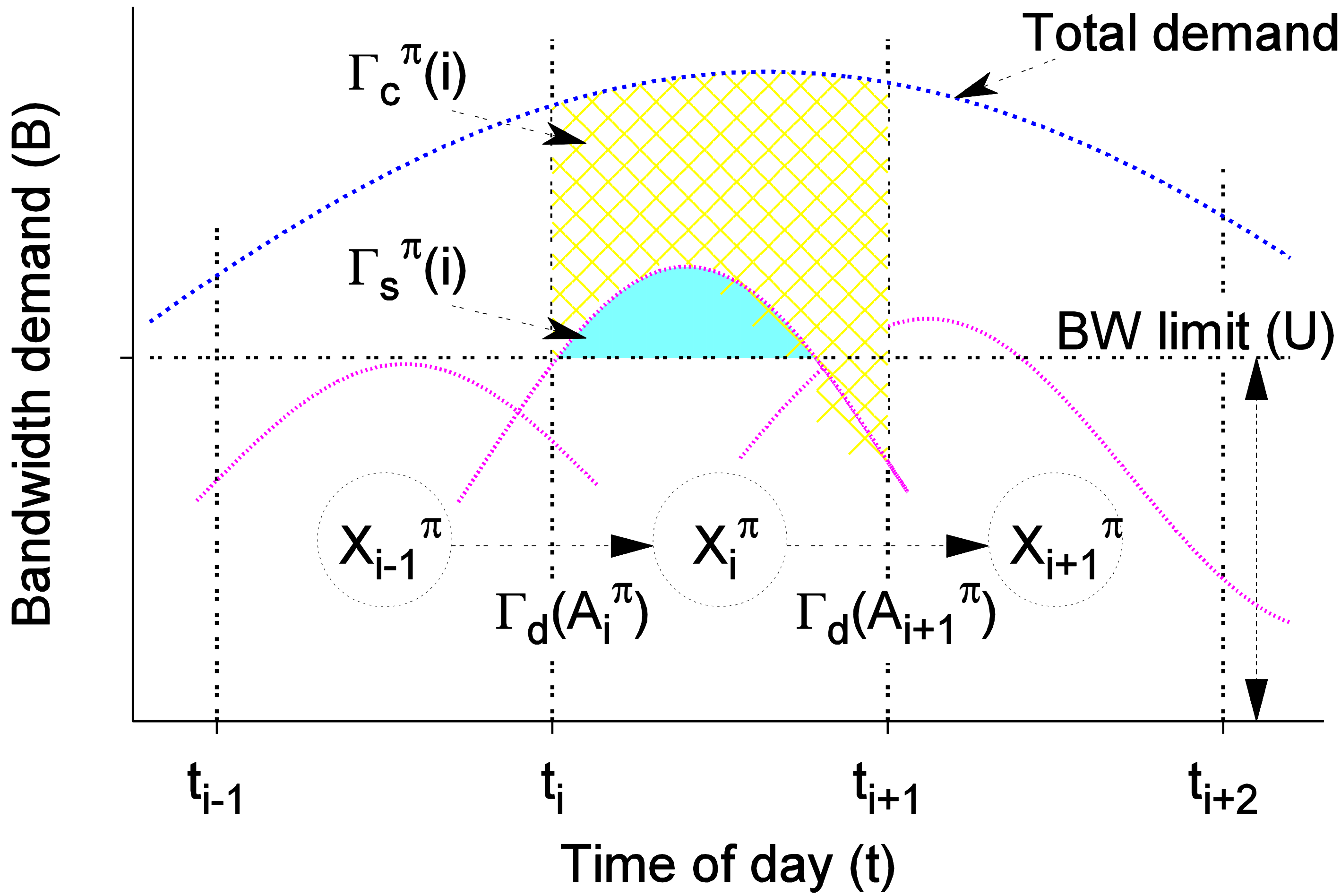
- How to get the best out of two worlds?
  - Improved workload models and predcition enables prefetching …
- Dynamic content allocation
  - Utilize capped bandwidth (and storage) as much as possible
  - Use elastic cloud-based services to serve “spillover”
-

Dynamic Content Allocation Problem

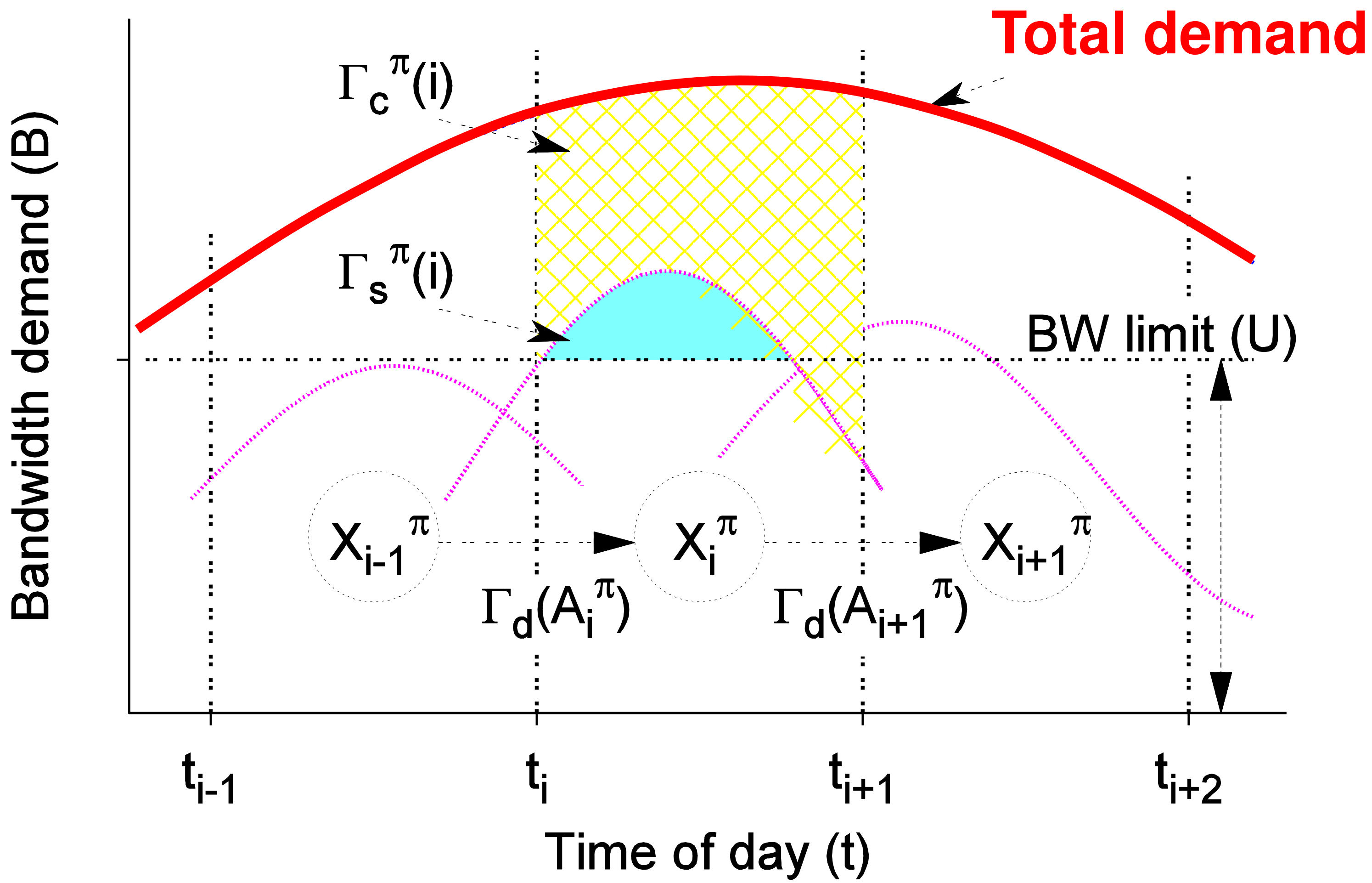


- Formulate as a finite horizon dynamic decision process problem
- Show discrete time decision process is good approximation
- Define exact solution as MILP
- Provide computationally feasible approximations (and prove properties about approximation ratios)
- Validate model and policies using traces from Spotify

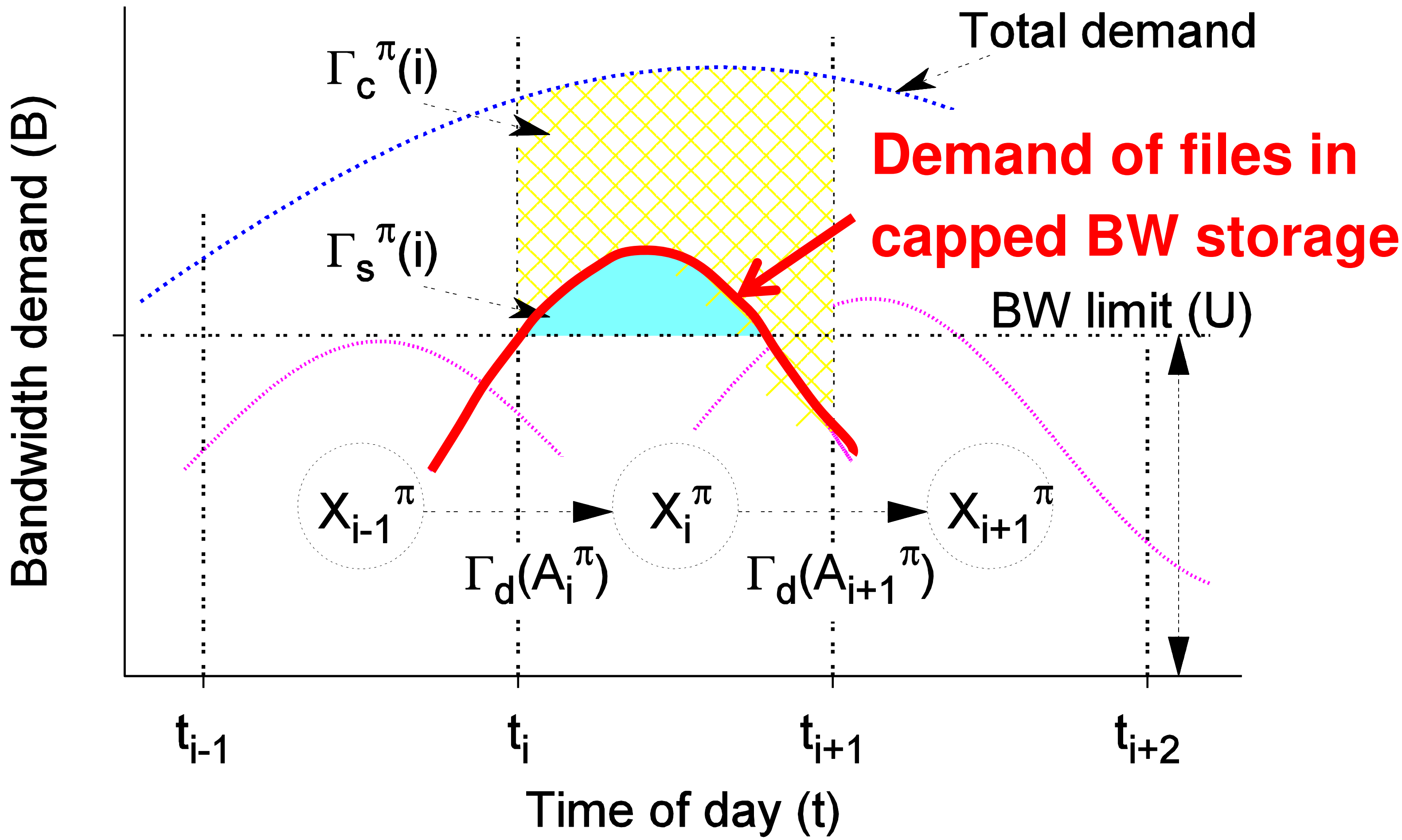
Cost minimization formulation



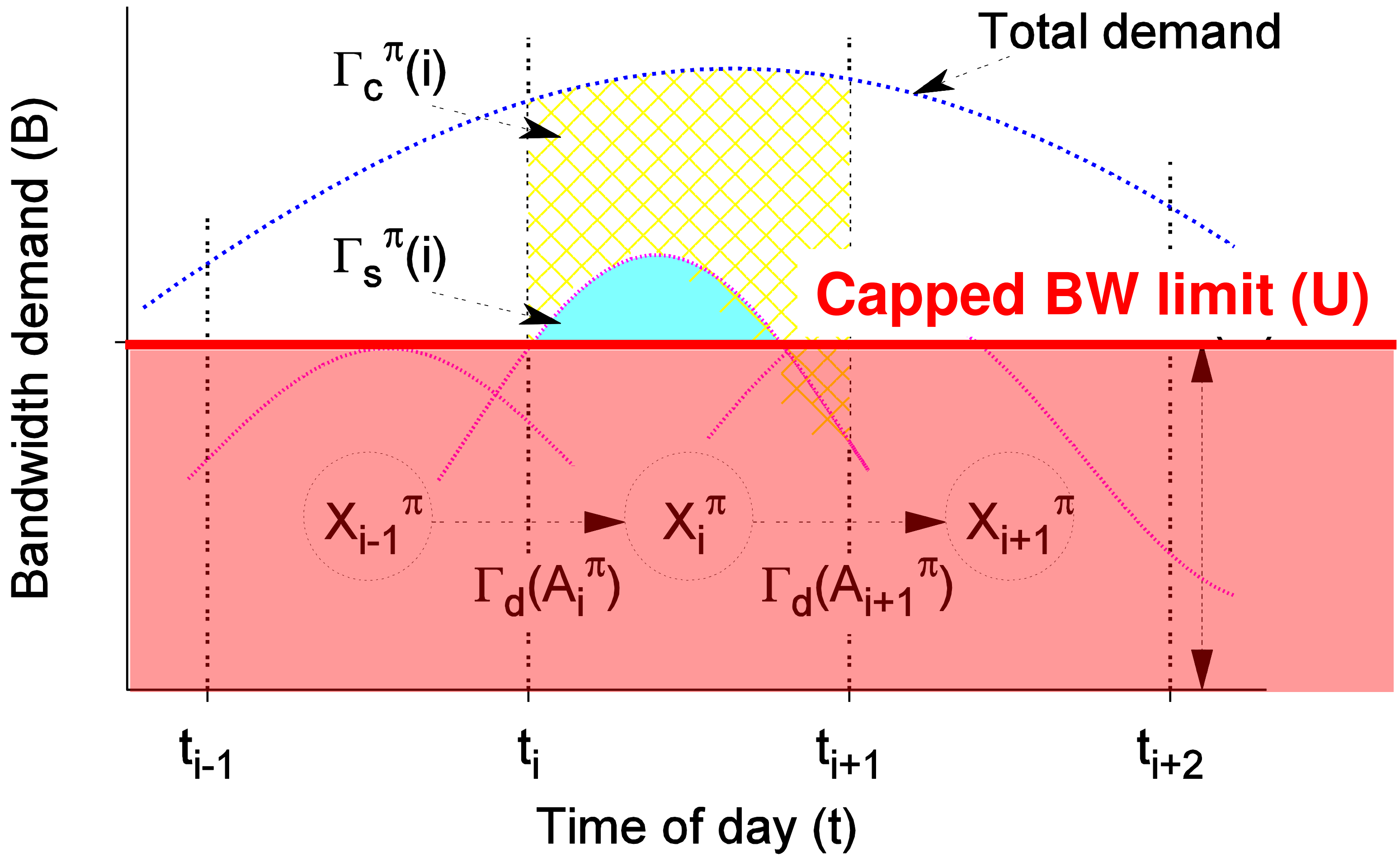
Cost minimization formulation



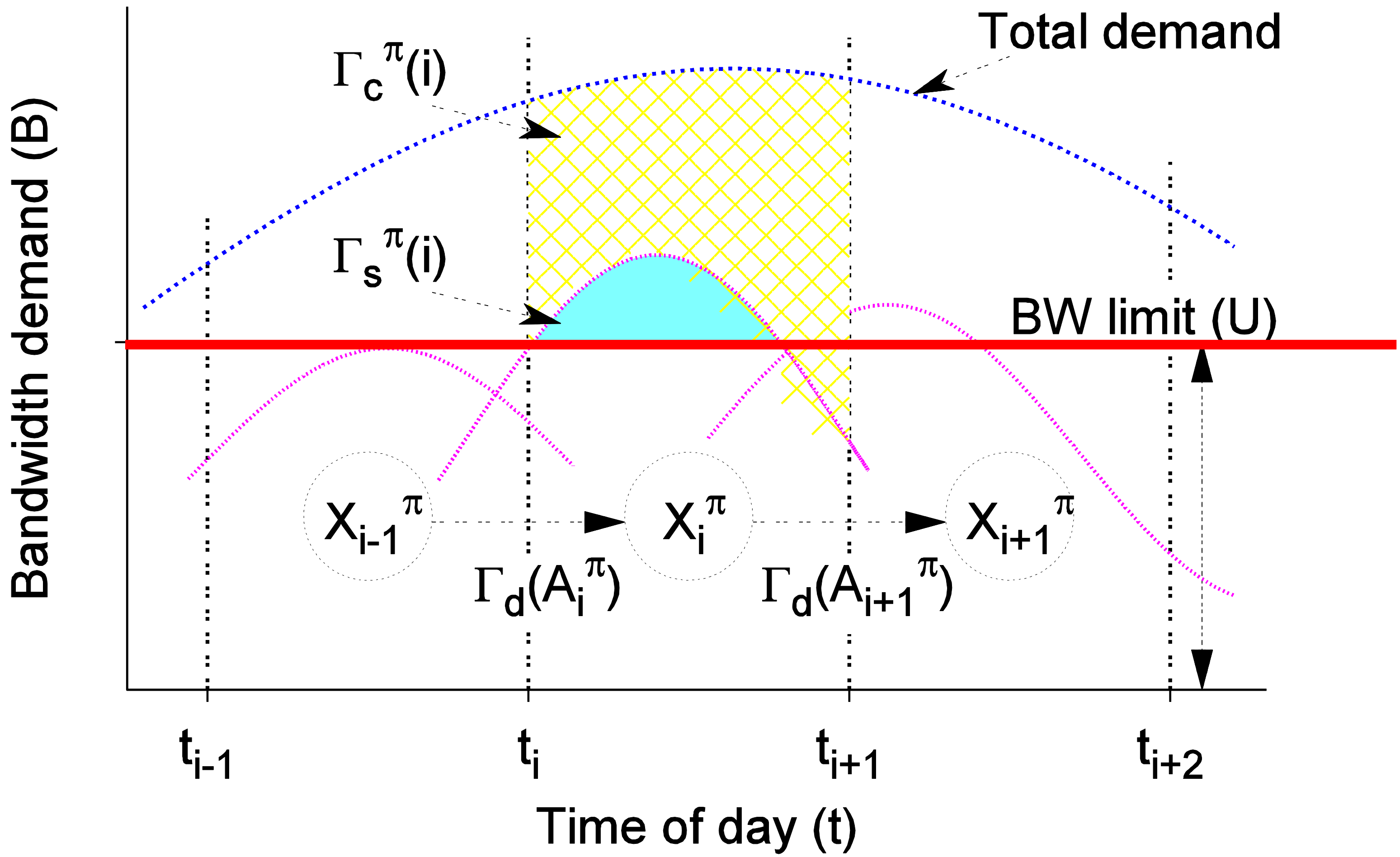
Cost minimization formulation



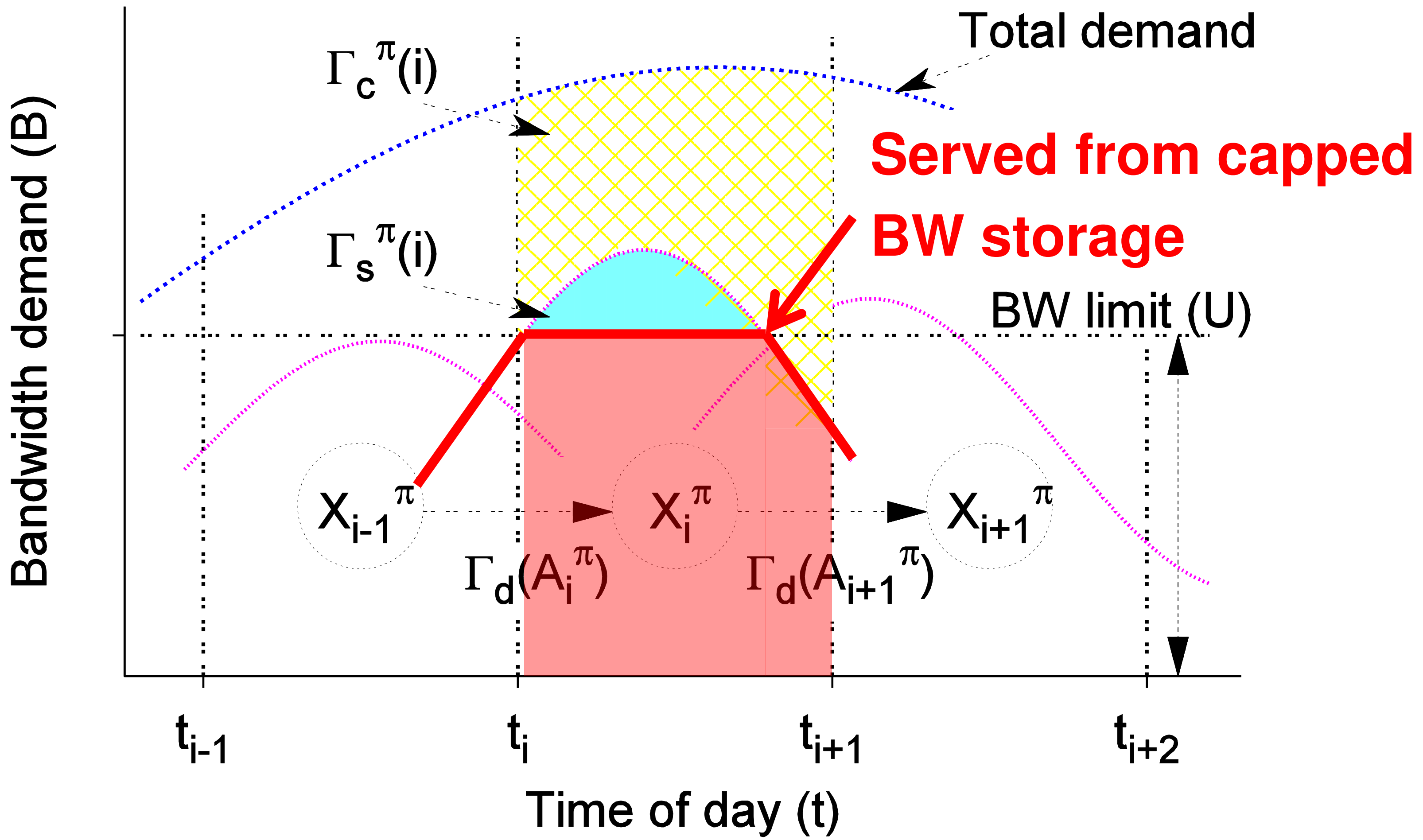
Cost minimization formulation



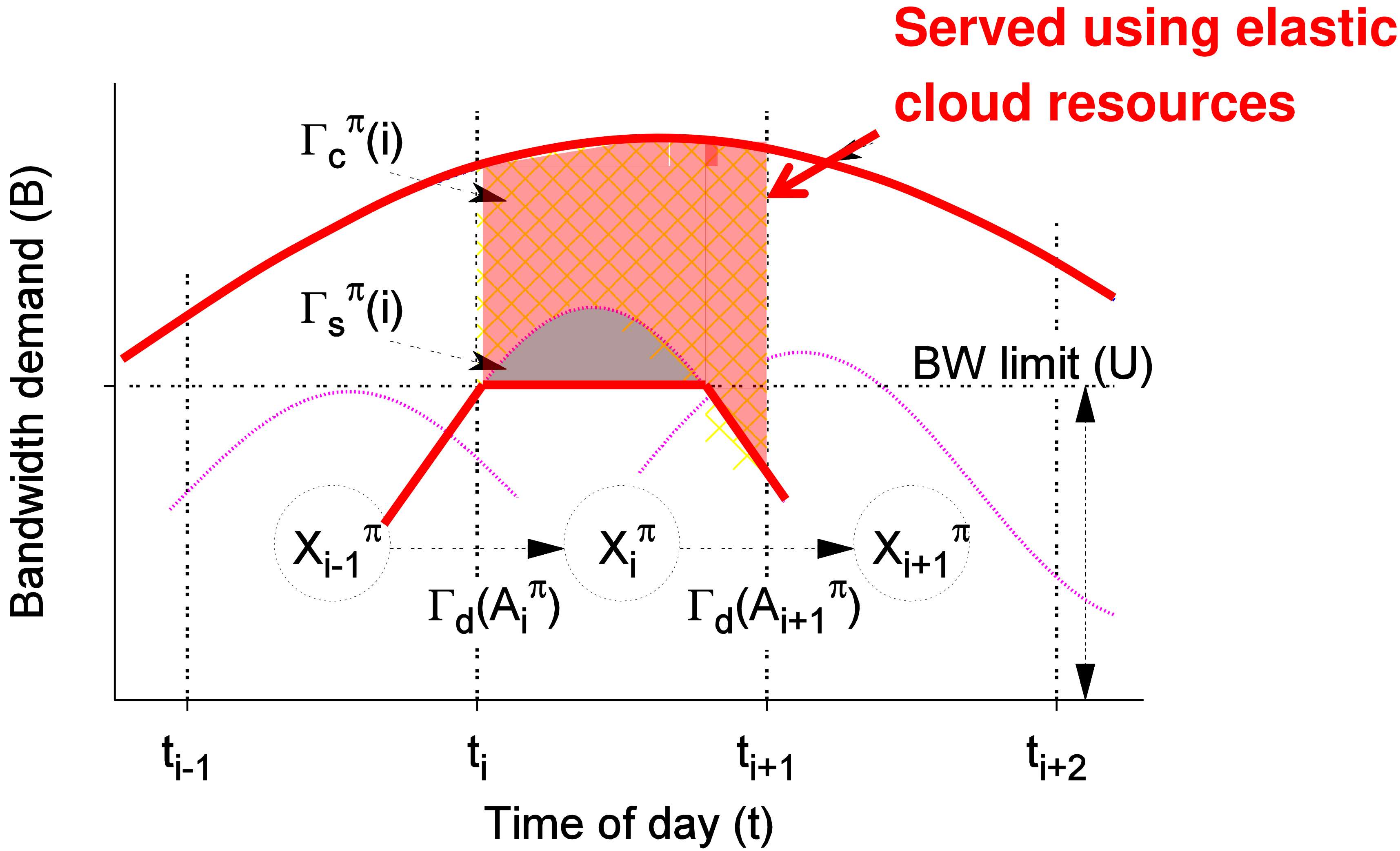
Cost minimization formulation



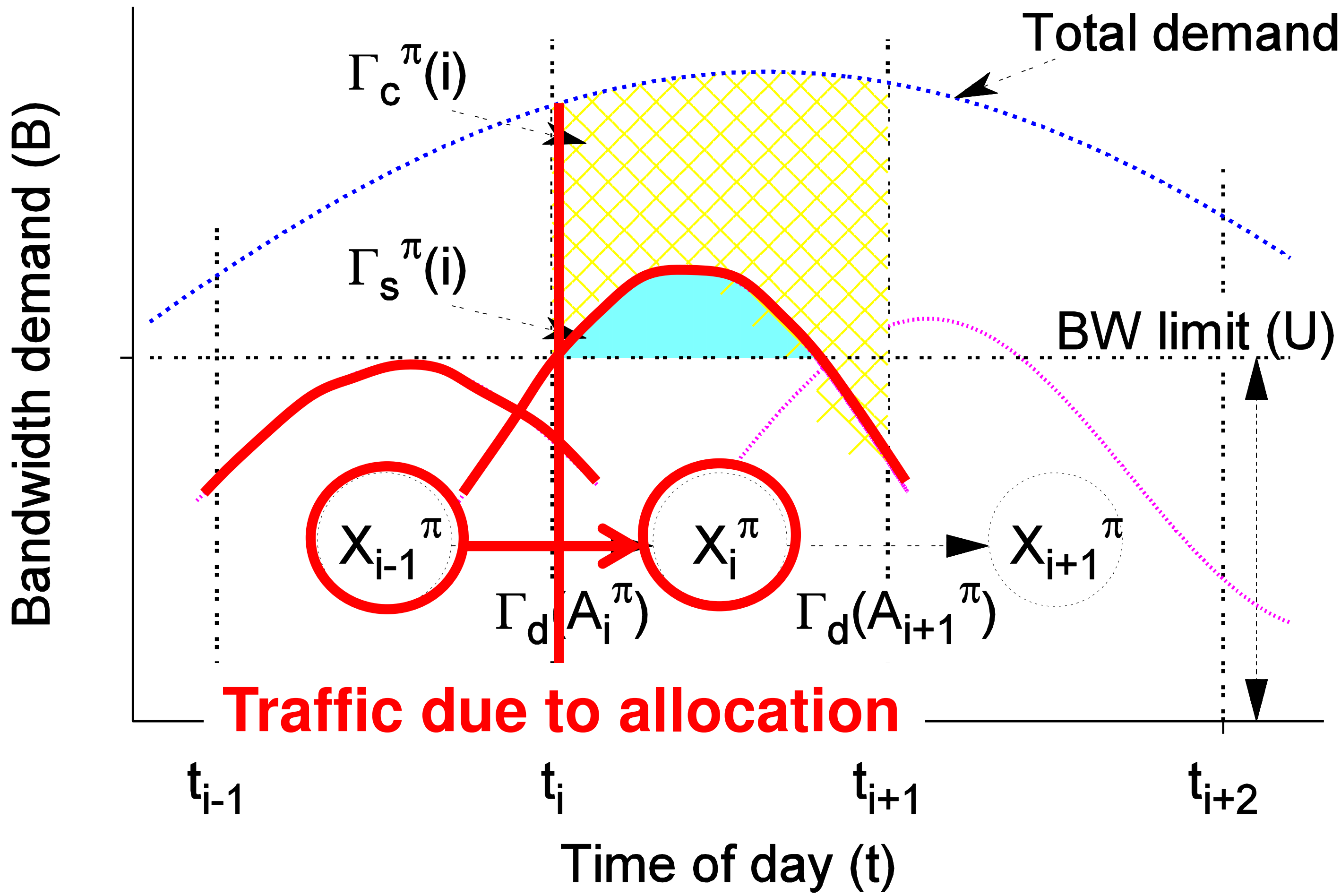
Cost minimization formulation



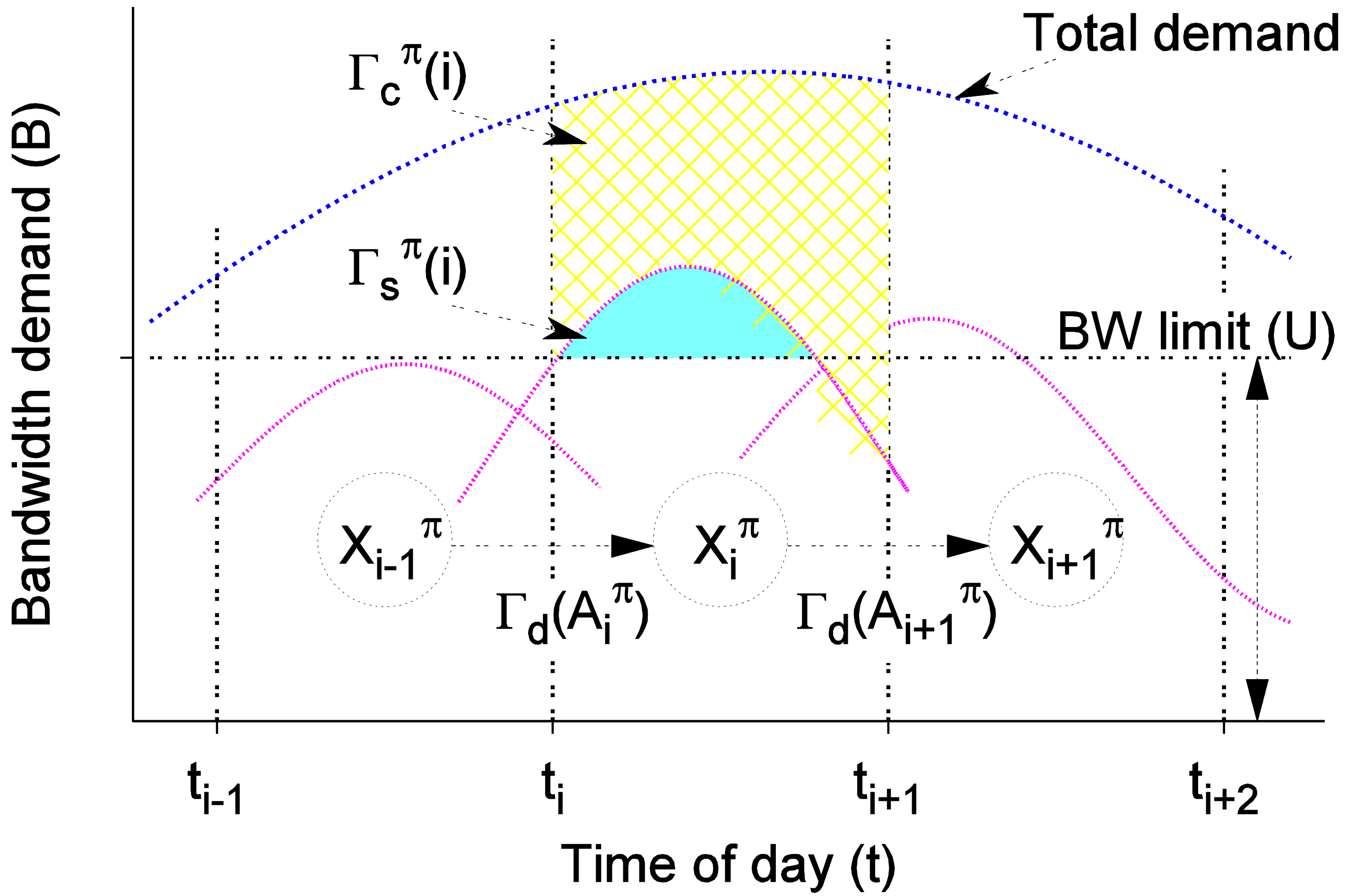
Cost minimization formulation



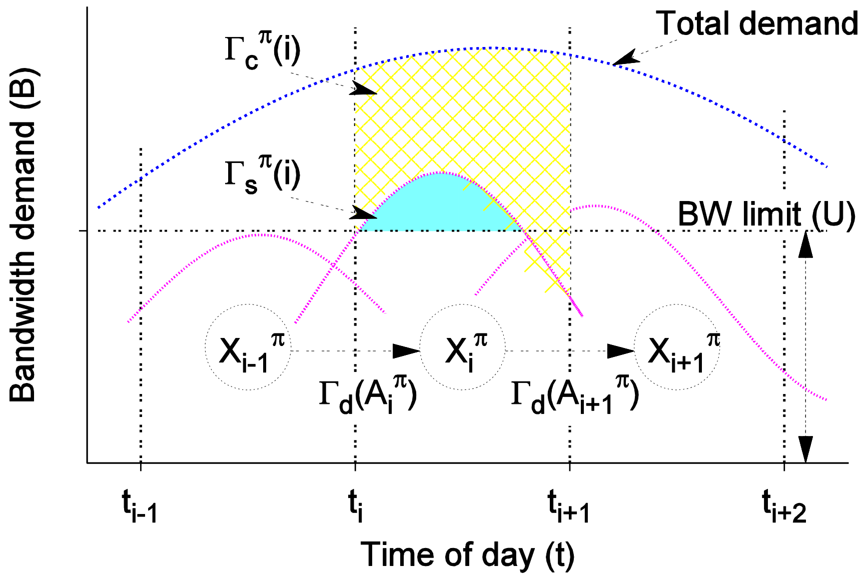
Cost minimization formulation



Cost minimization formulation



Cost minimization formulation



- Traffic of files only in cloud

$$\Gamma_c^{\pi}(i) = E\left[\int_{t_i^{\pi}}^{t_{i+1}^{\pi}} \sum_{f \notin \mathcal{X}_i^{\pi}} B_f(t)\right]$$

- Spillover traffic

$$\Gamma_s^{\pi}(i) = E\left[\int_{t_i^{\pi}}^{t_{i+1}^{\pi}} \left(\sum_{f \in \mathcal{X}_i^{\pi}} B_f(t) - U\right)^{+} dt\right]$$

- Traffic due to allocation

$$\Gamma_d^{\pi}(A_i^{\pi}) = \sum_{f \in A_i^{\pi}} L_f$$

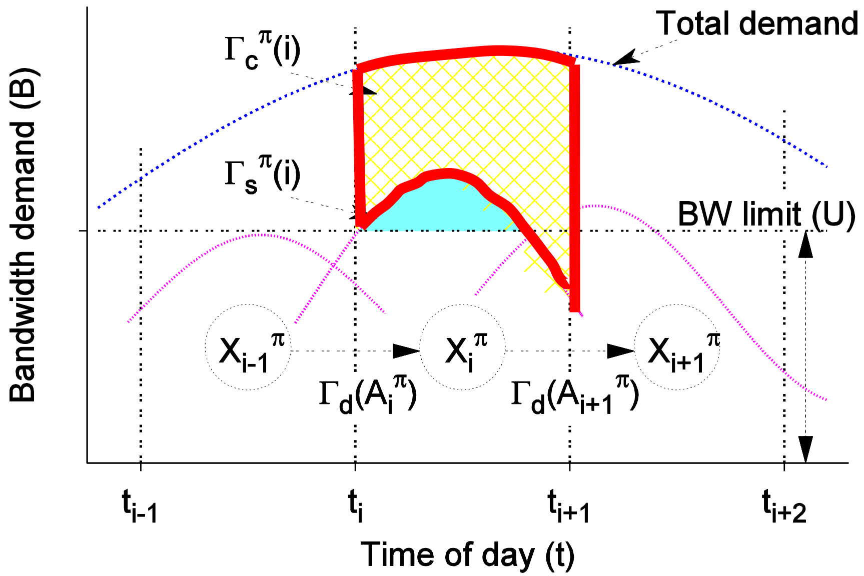
- Total expected cost

$$J^{\pi}(T, \mathcal{X}_0) = \gamma \times \sum_{i=0}^{I^{\pi}} \left\{\Gamma_d^{\pi}(A_i^{\pi}) + \Gamma_c^{\pi}(i) + \Gamma_s^{\pi}(i)\right\}$$

- Optimal policy

$$\pi^* = \arg\min_{\pi \in \Pi} J^{\pi}(T, \mathcal{X}_0)$$

Cost minimization formulation



- Traffic of files only in cloud

$$\Gamma_c^\pi(i) = E\left[\int_{t_i^\pi}^{t_{i+1}^\pi} \sum_{f \notin \mathcal{X}_i^\pi} B_f(t)\right]$$

- Spillover traffic

$$\Gamma_s^\pi(i) = E\left[\int_{t_i^\pi}^{t_{i+1}^\pi} \left(\sum_{f \in \mathcal{X}_i^\pi} B_f(t) - U\right)^+ dt\right]$$

- Traffic due to allocation

$$\Gamma_d^\pi(A_i^\pi) = \sum_{f \in A_i^\pi} L_f$$

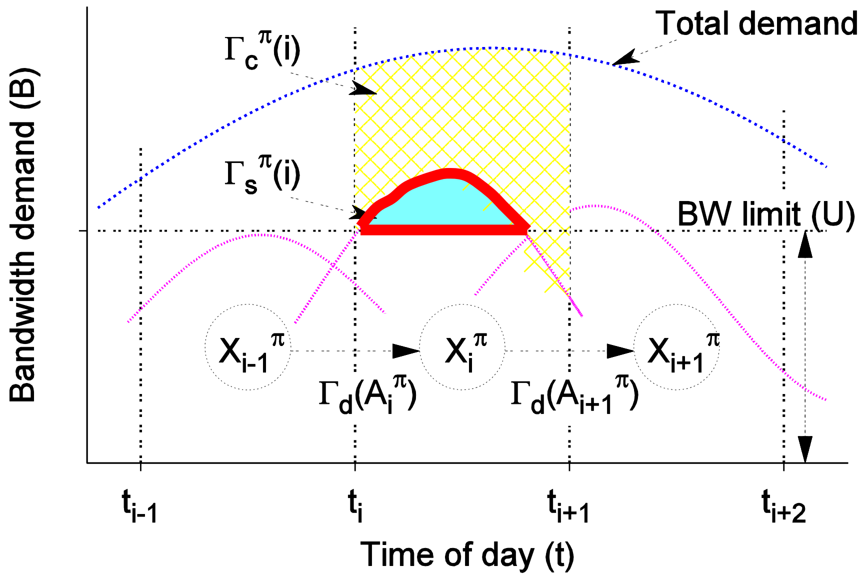
- Total expected cost

$$J^\pi(T, \mathcal{X}_0) = \gamma \times \sum_{i=0}^{I^\pi} \{\Gamma_d^\pi(A_i^\pi) + \Gamma_c^\pi(i) + \Gamma_s^\pi(i)\}$$

- Optimal policy

$$\pi^* = \arg\min_{\pi \in \Pi} J^\pi(T, \mathcal{X}_0)$$

Cost minimization formulation



- Traffic of files only in cloud

$$\Gamma_c^\pi(i) = E\left[\int_{t_i^\pi}^{t_{i+1}^\pi} \sum_{f \notin \mathcal{X}_i^\pi} B_f(t)\right]$$

- Spillover traffic

$$\Gamma_s^\pi(i) = E\left[\int_{t_i^\pi}^{t_{i+1}^\pi} \left(\sum_{f \in \mathcal{X}_i^\pi} B_f(t) - U\right)^+ dt\right]$$

- Traffic due to allocation

$$\Gamma_d^\pi(A_i^\pi) = \sum_{f \in A_i^\pi} L_f$$

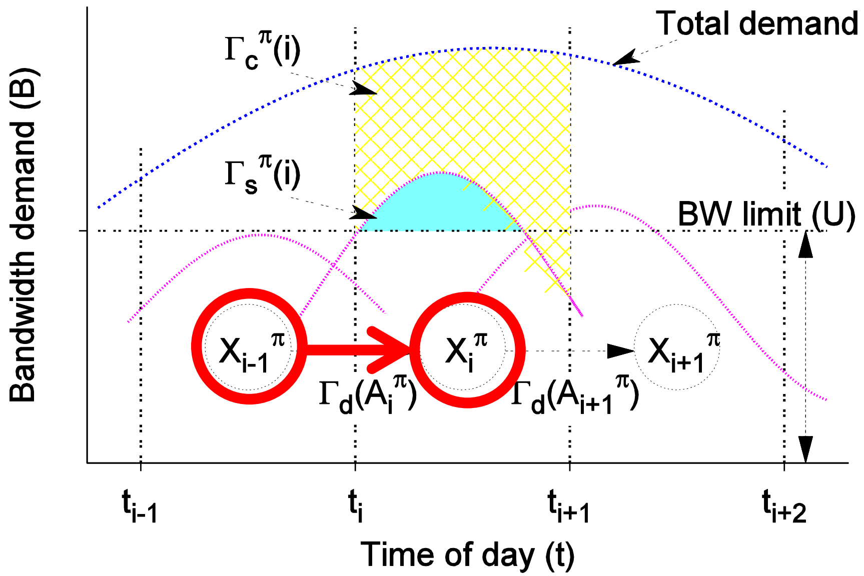
- Total expected cost

$$J^\pi(T, \mathcal{X}_0) = \gamma \times \sum_{i=0}^{I^\pi} \{\Gamma_d^\pi(A_i^\pi) + \Gamma_c^\pi(i) + \Gamma_s^\pi(i)\}$$

- Optimal policy

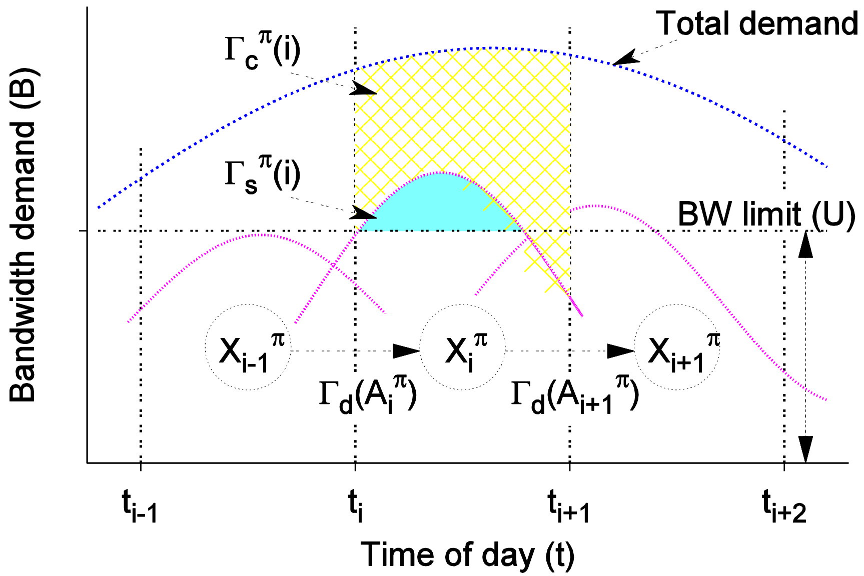
$$\pi^* = \arg\min_{\pi \in \Pi} J^\pi(T, \mathcal{X}_0)$$

Cost minimization formulation



- Traffic of files only in cloud
  $$\Gamma_c^{\pi}(i) = E\left[\int_{t_i^{\pi}}^{t_{i+1}^{\pi}} \sum_{f \notin \mathcal{X}_i^{\pi}} B_f(t)\right]$$
- Spillover traffic
  $$\Gamma_s^{\pi}(i) = E\left[\int_{t_i^{\pi}}^{t_{i+1}^{\pi}} \left(\sum_{f \in \mathcal{X}_i^{\pi}} B_f(t) - U\right)^{+} dt\right]$$
- Traffic due to allocation
  $$\Gamma_d^{\pi}(A_i^{\pi}) = \sum_{f \in A_i^{\pi}} L_f$$
- Total expected cost
  $$J^{\pi}(T, \mathcal{X}_0) = \gamma \times \sum_{i=0}^{I^{\pi}} \left\{\Gamma_d^{\pi}(A_i^{\pi}) + \Gamma_c^{\pi}(i) + \Gamma_s^{\pi}(i)\right\}$$
- Optimal policy
  $$\pi^* = \arg\min_{\pi \in \Pi} J^{\pi}(T, \mathcal{X}_0)$$

Cost minimization formulation



- Traffic of files only in cloud

$$\Gamma_c^\pi(i) = E\left[\int_{t_i^\pi}^{t_{i+1}^\pi} \sum_{f \notin \mathcal{X}_i^\pi} B_f(t)\right]$$

- Spillover traffic

$$\Gamma_s^\pi(i) = E\left[\int_{t_i^\pi}^{t_{i+1}^\pi} \left(\sum_{f \in \mathcal{X}_i^\pi} B_f(t) - U\right)^+ dt\right]$$

- Traffic due to allocation

$$\Gamma_d^\pi(A_i^\pi) = \sum_{f \in A_i^\pi} L_f$$

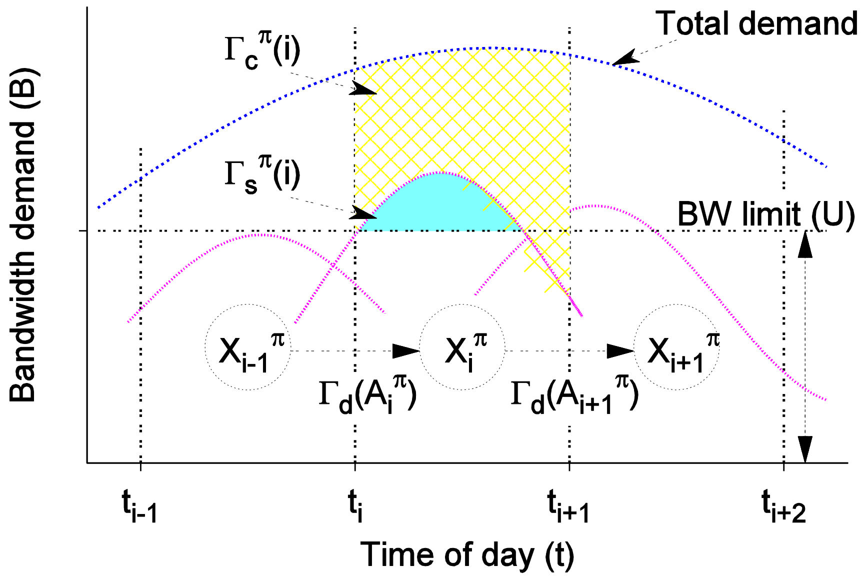
- Total expected cost

$$J^\pi(T, \mathcal{X}_0) = \gamma \times \sum_{i=0}^{I^\pi} \{\Gamma_d^\pi(A_i^\pi) + \Gamma_c^\pi(i) + \Gamma_s^\pi(i)\}$$

- Optimal policy

$$\pi^* = \arg\min_{\pi \in \Pi} J^\pi(T, \mathcal{X}_0)$$

Cost minimization formulation



- Traffic of files only in cloud

$$\Gamma_c^\pi(i) = E\left[\int_{t_i^\pi}^{t_{i+1}^\pi} \sum_{f \notin \mathcal{X}_i^\pi} B_f(t)\right]$$

- Spillover traffic

$$\Gamma_s^\pi(i) = E\left[\int_{t_i^\pi}^{t_{i+1}^\pi} \left(\sum_{f \in \mathcal{X}_i^\pi} B_f(t) - U\right)^+ dt\right]$$

- Traffic due to allocation

$$\Gamma_d^\pi(A_i^\pi) = \sum_{f \in A_i^\pi} L_f$$

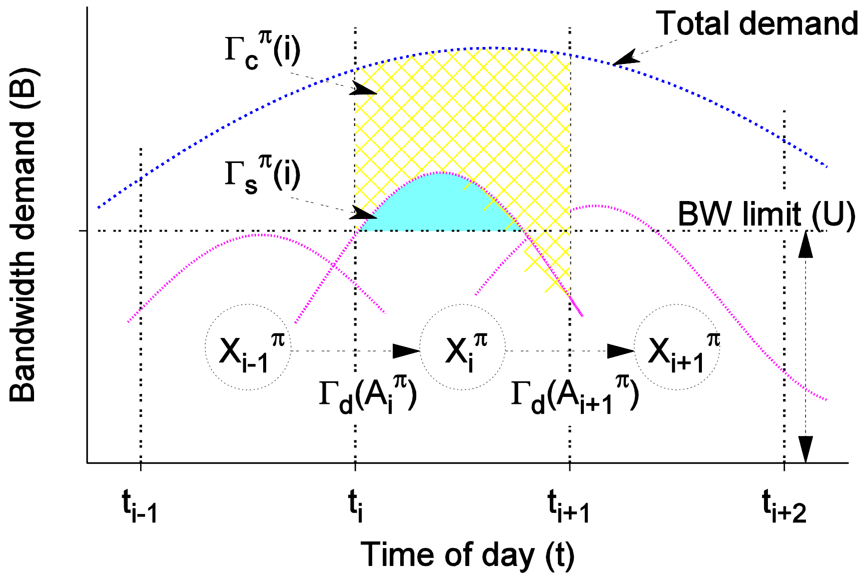
- Total expected cost

$$J^\pi(T, \mathcal{X}_0) = \gamma \times \sum_{i=0}^{I^\pi} \{\Gamma_d^\pi(A_i^\pi) + \Gamma_c^\pi(i) + \Gamma_s^\pi(i)\}$$

- Optimal policy

$$\pi^* = \arg\min_{\pi \in \Pi} J^\pi(T, \mathcal{X}_0)$$

Cost minimization formulation



- Traffic of files only in cloud

$$\Gamma_c^\pi(i) = E\left[\int_{t_i^\pi}^{t_{i+1}^\pi} \sum_{f \notin \mathcal{X}_i^\pi} B_f(t)\right]$$

- Spillover traffic

$$\Gamma_s^\pi(i) = E\left[\int_{t_i^\pi}^{t_{i+1}^\pi} \left(\sum_{f \in \mathcal{X}_i^\pi} B_f(t) - U\right)^+ dt\right]$$

- Traffic due to allocation

$$\Gamma_d^\pi(A_i^\pi) = \sum_{f \in A_i^\pi} L_f$$

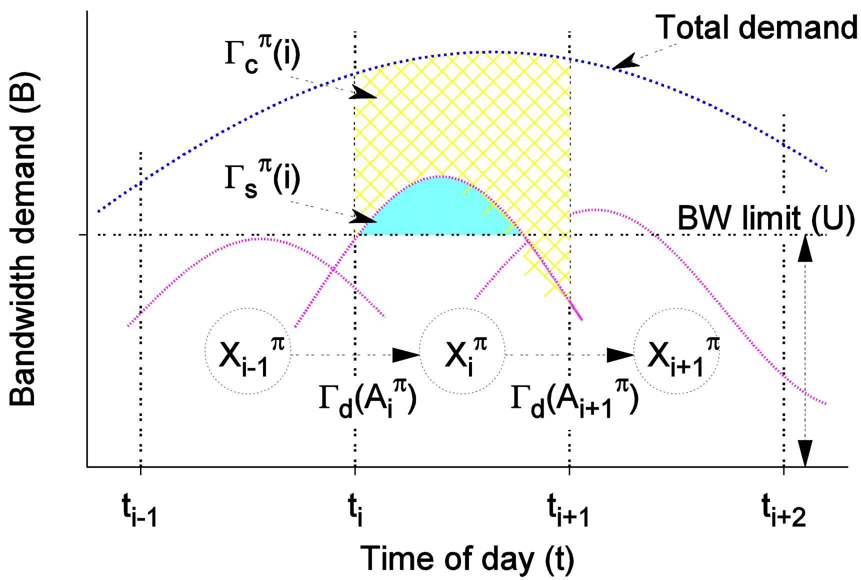
- Total expected cost

$$J^\pi(T, \mathcal{X}_0) = \gamma \times \sum_{i=0}^{I^\pi} \{\Gamma_d^\pi(A_i^\pi) + \Gamma_c^\pi(i) + \Gamma_s^\pi(i)\}$$

- Optimal policy

$$\pi^* = \arg\min_{\pi \in \Pi} J^\pi(T, \mathcal{X}_0)$$

Cost minimization formulation



- Traffic of files only in cloud

$$\Gamma_c^\pi(i) = E\left[\int_{t_i^\pi}^{t_{i+1}^\pi} \sum_{f \notin \mathcal{X}_i^\pi} B_f(t)\right]$$

- Spillover traffic

$$\Gamma_s^\pi(i) = E\left[\int_{t_i^\pi}^{t_{i+1}^\pi} \left(\sum_{f \in \mathcal{X}_i^\pi} B_f(t) - U\right)^+ dt\right]$$

- Traffic due to allocation

$$\Gamma_d^\pi(A_i^\pi) = \sum_{f \in A_i^\pi} L_f$$

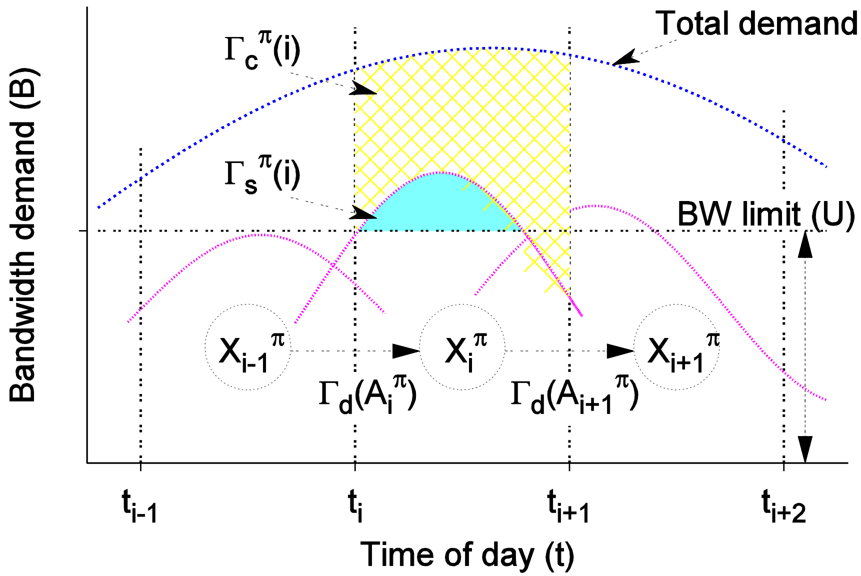
- Total expected cost

$$J^\pi(T, \mathcal{X}_0) = \gamma \times \sum_{i=0}^{I^\pi} \{\Gamma_d^\pi(A_i^\pi) + \Gamma_c^\pi(i) + \Gamma_s^\pi(i)\}$$

- Optimal policy

$$\pi^* = \arg\min_{\pi \in \Pi} J^\pi(T, \mathcal{X}_0)$$

Utilization maximization
~~Cost minimization formulation~~



- Traffic of files only in cloud

$$\Gamma_c^\pi(i) = E\left[\int_{t_i^\pi}^{t_{i+1}^\pi} \sum_{f \notin \mathcal{X}_i^\pi} B_f(t)\right]$$

- Spillover traffic

$$\Gamma_s^\pi(i) = E\left[\int_{t_i^\pi}^{t_{i+1}^\pi} \left(\sum_{f \in \mathcal{X}_i^\pi} B_f(t) - U\right)^+ dt\right]$$

- Traffic due to allocation

$$\Gamma_d^\pi(A_i^\pi) = \sum_{f \in A_i^\pi} L_f$$

- Total expected cost

$$J^\pi(T, \mathcal{X}_0) = \gamma \times \sum_{i=0}^{I^\pi} \left\{\Gamma_d^\pi(A_i^\pi) + \Gamma_c^\pi(i) + \Gamma_s^\pi(i)\right\}$$

- Optimal policy

$$\pi^* = \arg\min_{\pi \in \Pi} J^\pi(T, \mathcal{X}_0)$$

- Equivalent formulation

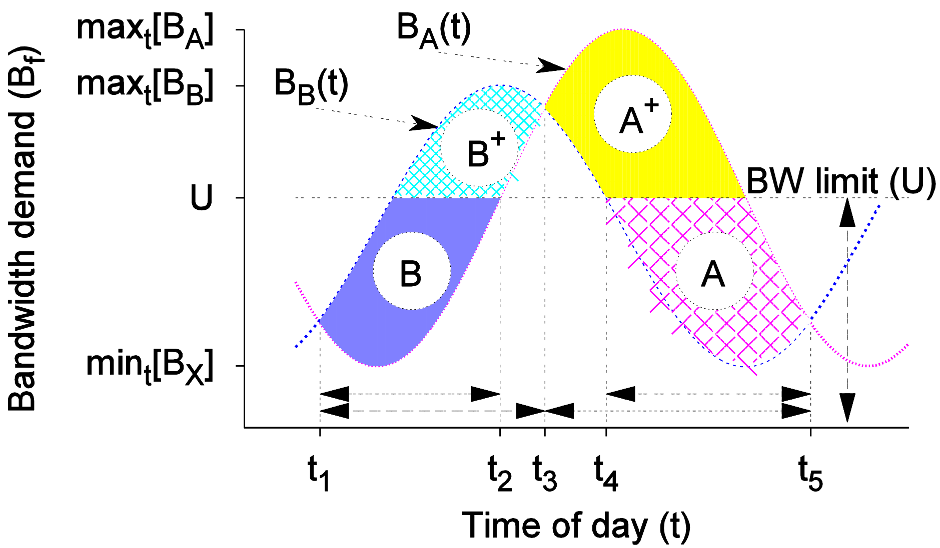
$$\overline{\Gamma}_s^\pi(i) = E\left[\int_{t_i^\pi}^{t_{i+1}^\pi} \min\left(U, \sum_{f \in \mathcal{X}_i^\pi} B_f(t)\right) dt\right]$$

$$U^\pi(T, \mathcal{X}_0) = \gamma \times \sum_{i=0}^{I^\pi} \left\{\overline{\Gamma}_s^\pi(i) - \Gamma_d^\pi(A_i^\pi)\right\}$$

Optimal policy $\pi^* = \arg\max_{\pi \in \Pi} U^\pi(T, \mathcal{X}_0)$

Utilization maximization

~~Cost minimization formulation~~



- Traffic of files only in cloud

$$\Gamma_c^\pi(i) = E\left[\int_{t_i^\pi}^{t_{i+1}^\pi} \sum_{f \notin \mathcal{X}_i^\pi} B_f(t)\right]$$

- Spillover traffic

$$\Gamma_s^\pi(i) = E\left[\int_{t_i^\pi}^{t_{i+1}^\pi} \left(\sum_{f \in \mathcal{X}_i^\pi} B_f(t) - U\right)^+ dt\right]$$

- Traffic due to allocation

$$\Gamma_d^\pi(A_i^\pi) = \sum_{f \in A_i^\pi} L_f$$

- Total expected cost

$$J^\pi(T, \mathcal{X}_0) = \gamma \times \sum_{i=0}^{I^\pi} \{\Gamma_d^\pi(A_i^\pi) + \Gamma_c^\pi(i) + \Gamma_s^\pi(i)\}$$

- Optimal policy

$$\pi^* = \arg\min_{\pi \in \Pi} J^\pi(T, \mathcal{X}_0)$$

- Equivalent formulation

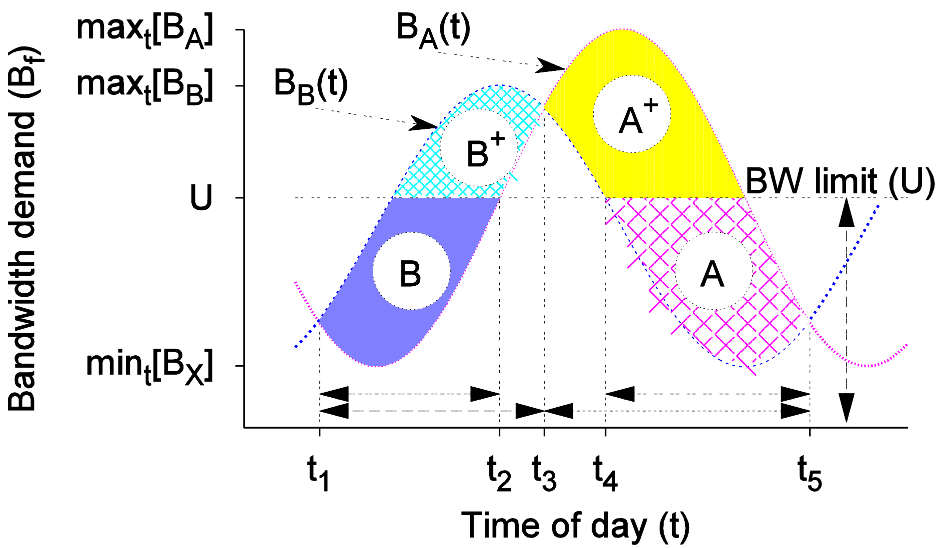
$$\overline{\Gamma}_s^\pi(i) = E\left[\int_{t_i^\pi}^{t_{i+1}^\pi} \min\left(U, \sum_{f \in \mathcal{X}_i^\pi} B_f(t)\right) dt\right]$$

$$U^\pi(T, \mathcal{X}_0) = \gamma \times \sum_{i=0}^{I^\pi} \left\{\overline{\Gamma}_s^\pi(i) - \Gamma_d^\pi(A_i^\pi)\right\}$$

Optimal policy $\pi^* = \arg\max_{\pi \in \Pi} U^\pi(T, \mathcal{X}_0)$

Utilization maximization

~~Cost minimization formulation~~



- Equivalent formulation

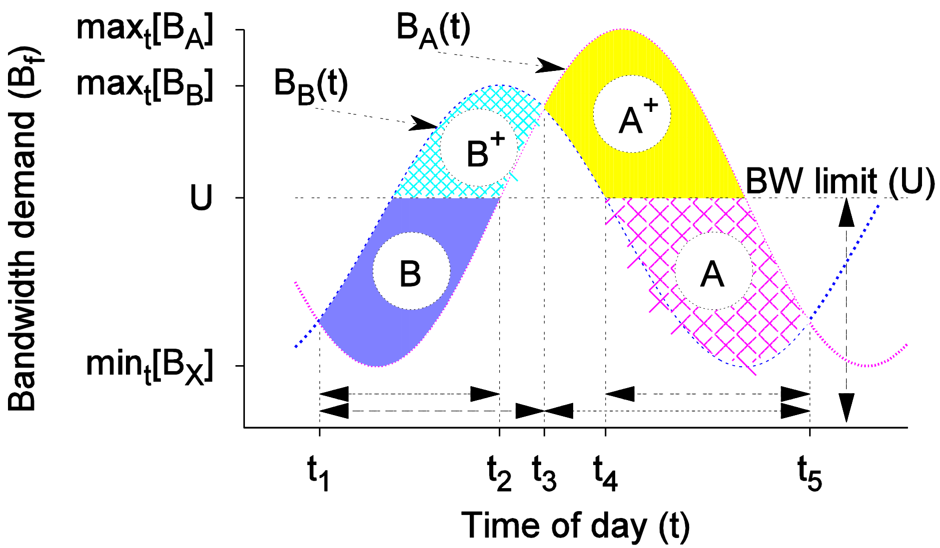
$$\overline{\Gamma}_s^{\pi}(i) = E\left[\int_{t_i^{\pi}}^{t_{i+1}^{\pi}} \min\left(U, \sum_{f \in \mathcal{X}_i^{\pi}} B_f(t)\right) dt\right]$$

$$U^{\pi}(T, \mathcal{X}_0) = \gamma \times \sum_{i=0}^{I^{\pi}} \left\{\overline{\Gamma}_s^{\pi}(i) - \Gamma_d^{\pi}(A_i^{\pi})\right\}$$

Optimal policy $\pi^* = \arg\max_{\pi \in \Pi} U^{\pi}(T, \mathcal{X}_0)$

Utilization maximization

~~Cost minimization formulation~~



Two file example

- Equivalent formulation

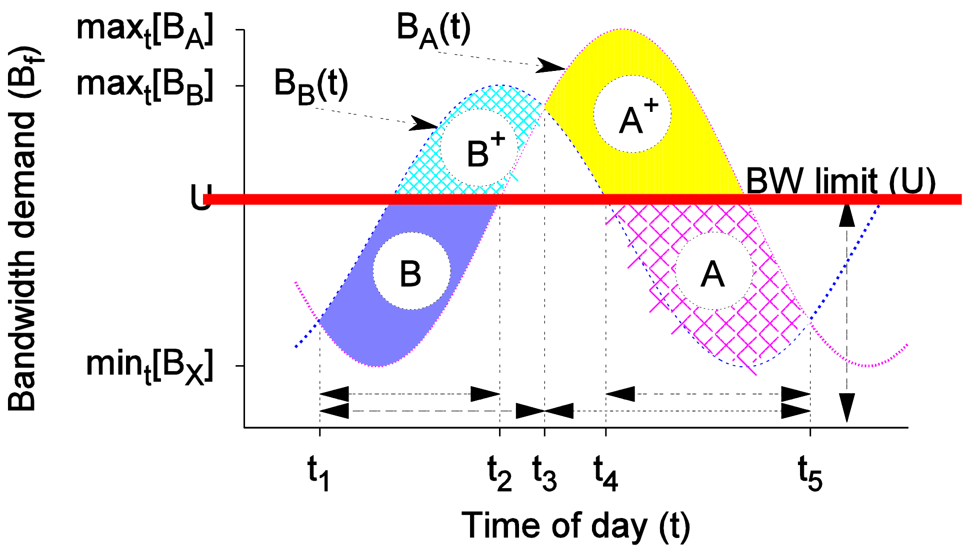
$$\overline{\Gamma}^{\pi}_s(i) = E\left[\int_{t_i^\pi}^{t_{i+1}^\pi} \min\left(U, \sum_{f\in\mathcal{X}_i^\pi} B_f(t)\right) dt\right]$$

$$U^\pi(T,\mathcal{X}_0) = \gamma \times \sum_{i=0}^{I^\pi}\left\{\overline{\Gamma}^\pi_s(i) - \Gamma^\pi_d(A_i^\pi)\right\}$$

Optimal policy $\pi^* = \arg\max_{\pi\in\Pi} U^\pi(T,\mathcal{X}_0)$

Utilization maximization

~~Cost minimization formulation~~



Two file example

- Equivalent formulation

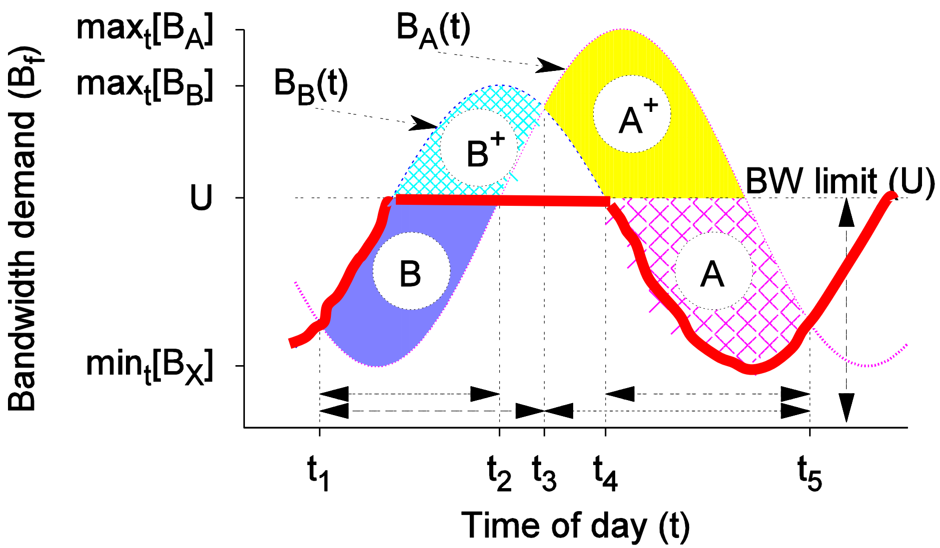
$$\overline{\Gamma}^{\pi}_{s}(i) = E\left[\int_{t_i^{\pi}}^{t_{i+1}^{\pi}} \min\left(U, \sum_{f \in \mathcal{X}_i^{\pi}} B_f(t)\right) dt\right]$$

$$U^{\pi}(T, \mathcal{X}_0) = \gamma \times \sum_{i=0}^{I^{\pi}} \left\{\overline{\Gamma}^{\pi}_{s}(i) - \Gamma^{\pi}_{d}(A^{\pi}_{i})\right\}$$

Optimal policy $\pi^* = \arg\max_{\pi \in \Pi} U^{\pi}(T, \mathcal{X}_0)$

Utilization maximization

~~Cost minimization formulation~~



Two file example

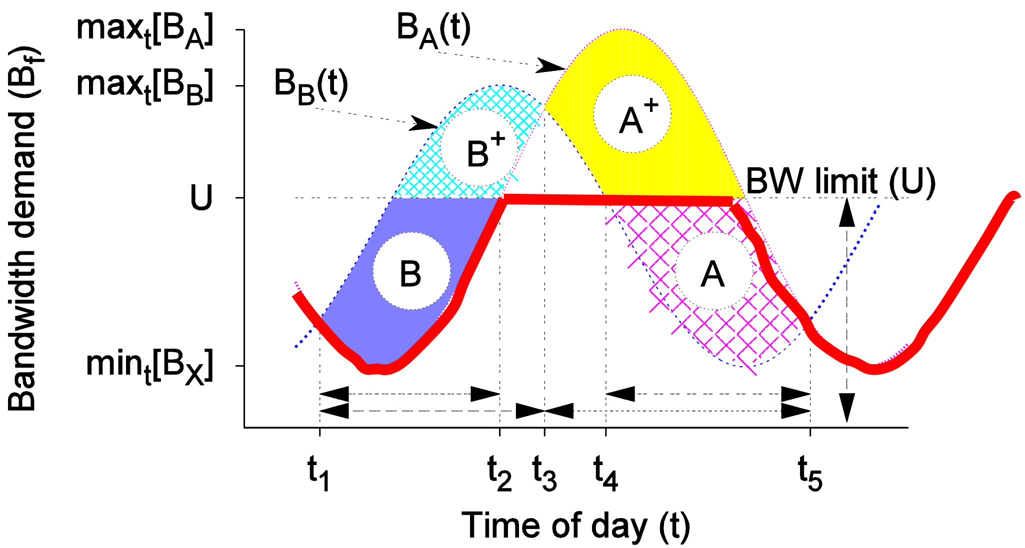
- Equivalent formulation

$$\overline{\Gamma}^{\pi}_{s}(i) = E\left[\int_{t_i^{\pi}}^{t_{i+1}^{\pi}} \min\left(U, \sum_{f\in\mathcal{X}_i^{\pi}} B_f(t)\right) dt\right]$$

$$U^{\pi}(T,\mathcal{X}_0) = \gamma \times \sum_{i=0}^{I^{\pi}} \left\{\overline{\Gamma}^{\pi}_{s}(i) - \Gamma^{\pi}_{d}(A_i^{\pi})\right\}$$

Optimal policy $\pi^* = \arg\max_{\pi\in\Pi} U^{\pi}(T,\mathcal{X}_0)$

Utilization maximization
~~Cost minimization formulation~~



Two file example

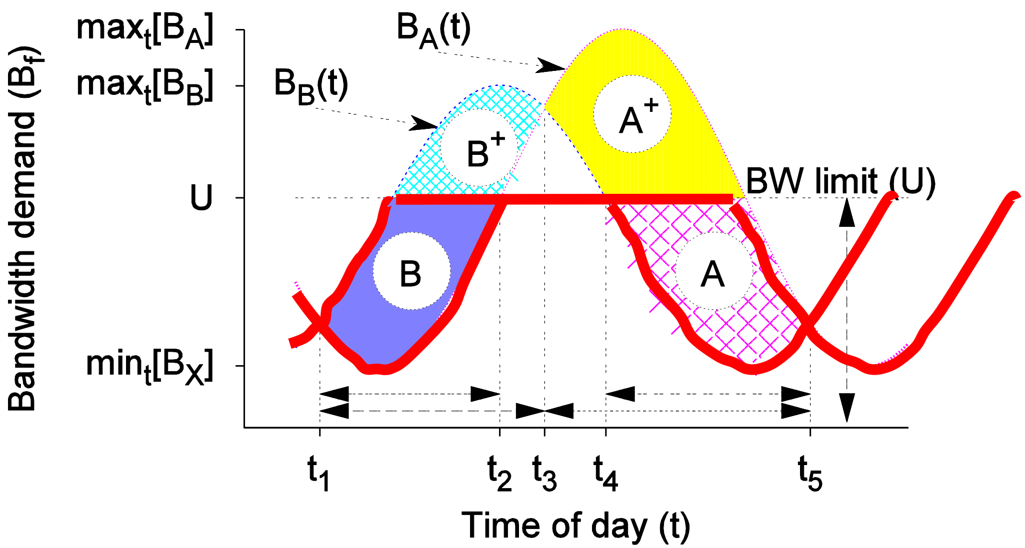
- Equivalent formulation

$$\overline{\Gamma}^{\pi}_{s}(i) = E\left[\int_{t_i^{\pi}}^{t_{i+1}^{\pi}} \min\left(U, \sum_{f\in\mathcal{X}_i^{\pi}} B_f(t)\right) dt\right]$$

$$U^{\pi}(T,\mathcal{X}_0) = \gamma \times \sum_{i=0}^{I^{\pi}} \left\{\overline{\Gamma}^{\pi}_{s}(i) - \Gamma^{\pi}_{d}(A_i^{\pi})\right\}$$

Optimal policy $\pi^* = \arg\max_{\pi\in\Pi} U^{\pi}(T,\mathcal{X}_0)$

Utilization maximization
~~Cost minimization formulation~~



Two file example

- Equivalent formulation

$$\overline{\Gamma}_s^\pi(i) = E\left[\int_{t_i^\pi}^{t_{i+1}^\pi} \min\left(U, \sum_{f\in\mathcal{X}_i^\pi} B_f(t)\right) dt\right]$$

$$U^\pi(T,\mathcal{X}_0) = \gamma \times \sum_{i=0}^{I^\pi}\left\{\overline{\Gamma}_s^\pi(i) - \Gamma_d^\pi(A_i^\pi)\right\}$$

Optimal policy $\pi^* = \arg\max_{\pi\in\Pi} U^\pi(T,\mathcal{X}_0)$

Utilization maximization
~~Cost minimization formulation~~

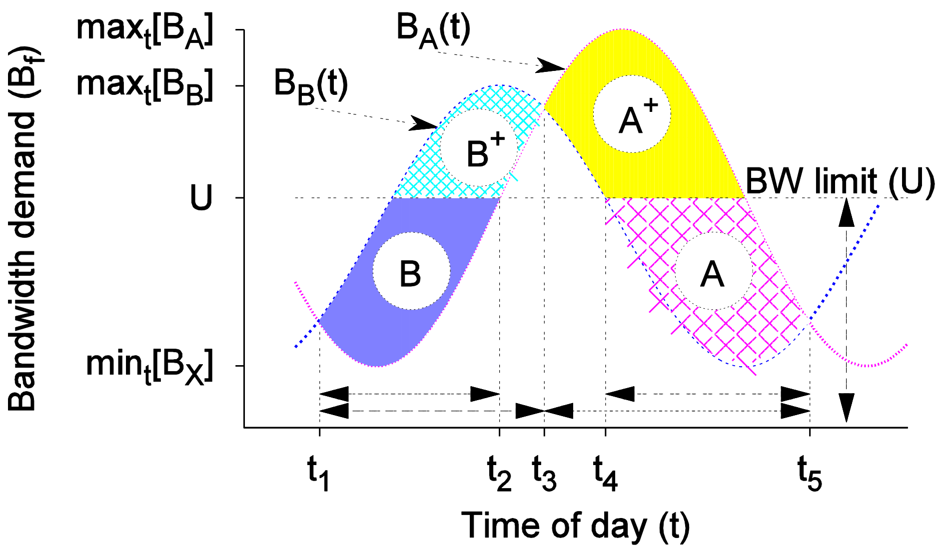
- Equivalent formulation

$$\overline{\Gamma}_s^{\pi}(i) = E\left[\int_{t_i^{\pi}}^{t_{i+1}^{\pi}} \min\left(U, \sum_{f \in \mathcal{X}_i^{\pi}} B_f(t)\right) dt\right]$$

$$U^{\pi}(T, \mathcal{X}_0) = \gamma \times \sum_{i=0}^{I^{\pi}} \left\{\overline{\Gamma}_s^{\pi}(i) - \Gamma_d^{\pi}(A_i^{\pi})\right\}$$

Optimal policy $\pi^* = \arg\max_{\pi \in \Pi} U^{\pi}(T, \mathcal{X}_0)$

Discrete-time Decision Problem



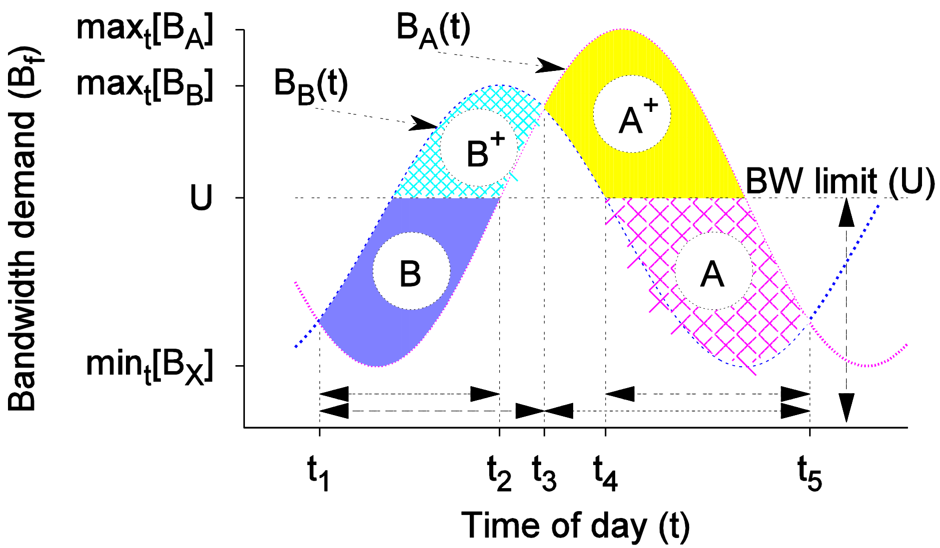
- Equivalent formulation

$$\overline{\Gamma}_s^\pi(i) = E\left[\int_{t_i^\pi}^{t_{i+1}^\pi} \min\left(U, \sum_{f\in\mathcal{X}_i^\pi} B_f(t)\right) dt\right]$$

$$U^\pi(T,\mathcal{X}_0) = \gamma \times \sum_{i=0}^{I^\pi} \left\{\overline{\Gamma}_s^\pi(i) - \Gamma_d^\pi(A_i^\pi)\right\}$$

Optimal policy $\pi^* = \arg\max_{\pi\in\Pi} U^\pi(T,\mathcal{X}_0)$

Discrete-time Decision Problem



- Approximation

$$\sum_{f\in\mathcal{X}(t)} B_f(t) \approx \sum_{f\in\mathcal{X}_i} \overline{B}_f^i \text{ for } t_i \le t < t_{i+1}$$

$P(\sum_{f\in\mathcal{X}} B_f(t) \le U)$ *decrease exponentially*

- Finite horizon decision

$$U^{\pi^*}([t_i, t_{I+1}], \mathcal{X}_{i-1}) = \max_{\mathcal{X}_i} \{\overline{\Gamma}_s(i) - \Gamma_d(A_i) + U^{\pi^*}([t_{i+1}, t_{I+1}], \mathcal{X}_i)\}$$

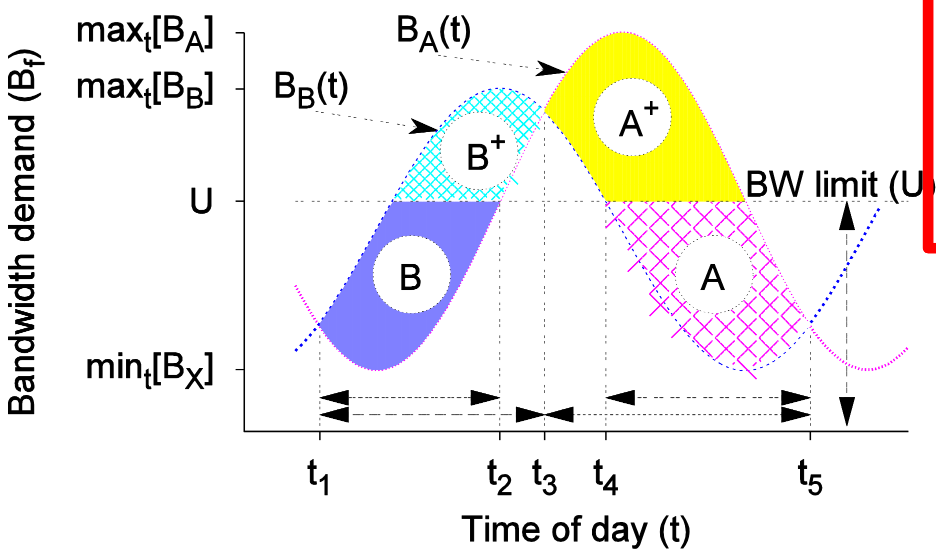
- Equivalent formulation

$$\overline{\Gamma}_s^\pi(i) = E\left[\int_{t_i^\pi}^{t_{i+1}^\pi} \min\left(U, \sum_{f\in\mathcal{X}_i^\pi} B_f(t)\right) dt\right]$$

$$U^\pi(T, \mathcal{X}_0) = \gamma \times \sum_{i=0}^{I^\pi} \left\{\overline{\Gamma}_s^\pi(i) - \Gamma_d^\pi(A_i^\pi)\right\}$$

Optimal policy $\pi^* = \arg\max_{\pi\in\Pi} U^\pi(T, \mathcal{X}_0)$

Discrete-time Decision Problem



- Approximation

$$\sum_{f \in \mathcal{X}(t)} B_f(t) \approx \sum_{f \in \mathcal{X}_i} \overline{B}_f^i \text{ for } t_i \leq t < t_{i+1}$$

$P(\sum_{f \in \mathcal{X}} B_f(t) \leq U)$ *decrease exponentially*

- Finite horizon decision

$$U^{\pi^*}([t_i, t_{I+1}], \mathcal{X}_{i-1}) = \max_{\mathcal{X}_i} \{\overline{\Gamma}_s(i) - \Gamma_d(A_i) + U^{\pi^*}([t_{i+1}, t_{I+1}], \mathcal{X}_i)\}$$

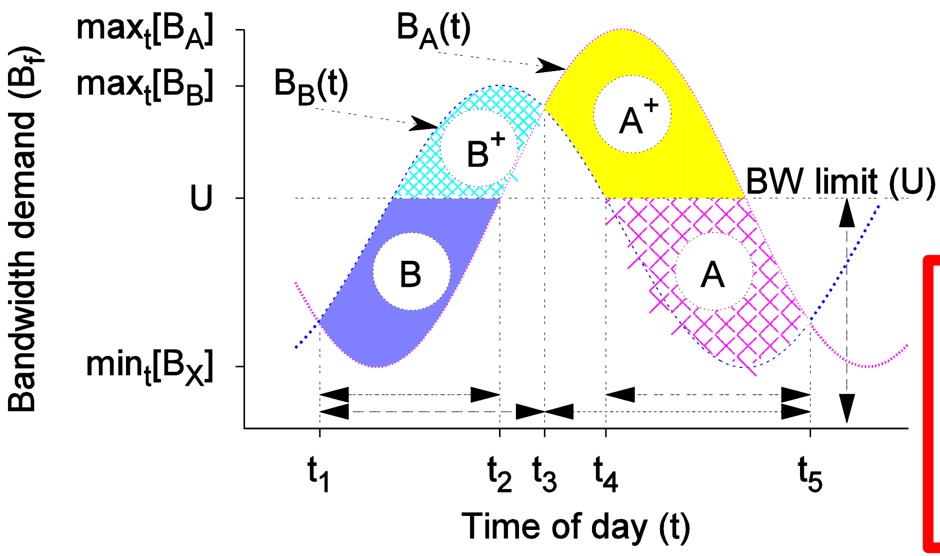
- Equivalent formulation

$$\overline{\Gamma}_s^\pi(i) = E\left[\int_{t_i^\pi}^{t_{i+1}^\pi} \min\left(U, \sum_{f \in \mathcal{X}_i^\pi} B_f(t)\right) dt\right]$$

$$U^\pi(T, \mathcal{X}_0) = \gamma \times \sum_{i=0}^{I^\pi} \left\{\overline{\Gamma}_s^\pi(i) - \Gamma_d^\pi(A_i^\pi)\right\}$$

Optimal policy $\pi^* = \arg\max_{\pi \in \Pi} U^\pi(T, \mathcal{X}_0)$

Discrete-time Decision Problem



- Approximation

$$\sum_{f\in\mathcal{X}(t)} B_f(t) \approx \sum_{f\in\mathcal{X}_i} \overline{B}_f^i \text{ for } t_i \le t < t_{i+1}$$

$P(\sum_{f\in\mathcal{X}} B_f(t) \le U)$ *decrease exponentially*

- Finite horizon decision

$$U^{\pi^*}([t_i, t_{I+1}], \mathcal{X}_{i-1}) =$$
$$\max_{\mathcal{X}_i} \{\overline{\Gamma}_s(i) - \Gamma_d(A_i) + U^{\pi^*}([t_{i+1}, t_{I+1}], \mathcal{X}_i)\}$$

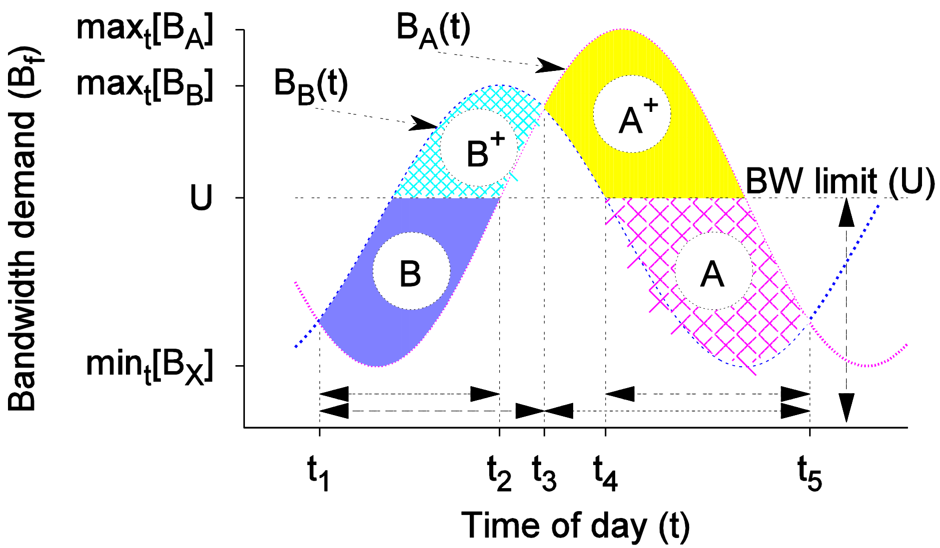
- Equivalent formulation

$$\overline{\Gamma}_s^\pi(i) = E\left[\int_{t_i^\pi}^{t_{i+1}^\pi} \min\left(U, \sum_{f\in\mathcal{X}_i^\pi} B_f(t)\right) dt\right]$$

$$U^\pi(T, \mathcal{X}_0) = \gamma \times \sum_{i=0}^{I^\pi} \left\{\overline{\Gamma}_s^\pi(i) - \Gamma_d^\pi(A_i^\pi)\right\}$$

Optimal policy $\pi^* = \arg\max_{\pi\in\Pi} U^\pi(T, \mathcal{X}_0)$

Discrete-time Decision Problem



- Approximation

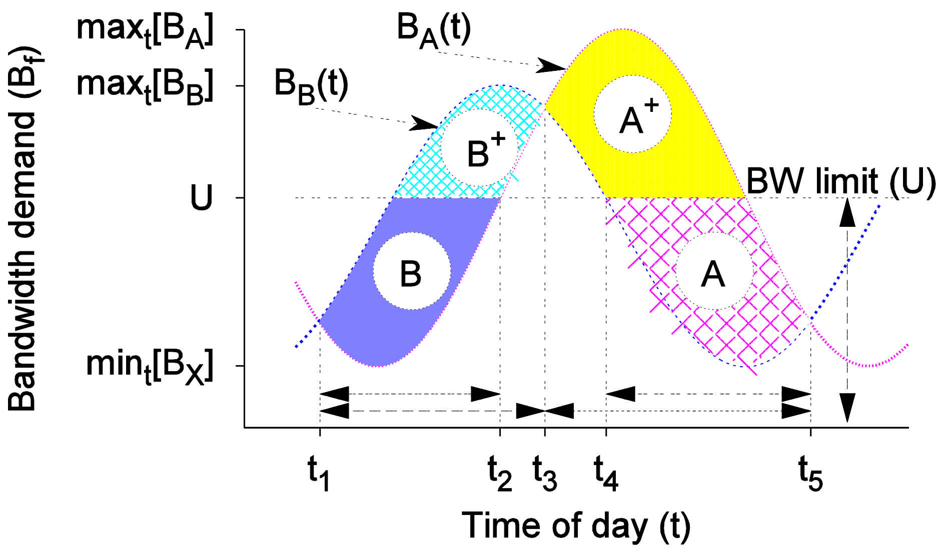
$$\sum_{f\in\mathcal{X}(t)} B_f(t) \approx \sum_{f\in\mathcal{X}_i} \overline{B}_f^i \text{ for } t_i \le t < t_{i+1}$$

$P(\sum_{f\in\mathcal{X}} B_f(t) \le U)$ *decrease exponentially*

- Finite horizon decision

$$U^{\pi^*}([t_i, t_{I+1}], \mathcal{X}_{i-1}) = \max_{\mathcal{X}_i} \{\overline{\Gamma}_s(i) - \Gamma_d(A_i) + U^{\pi^*}([t_{i+1}, t_{I+1}], \mathcal{X}_i)\}$$

Discrete-time Decision Problem



- Approximation

$$\sum_{f\in\mathcal{X}(t)} B_f(t) \approx \sum_{f\in\mathcal{X}_i} \overline{B}^i_f \text{ for } t_i \le t < t_{i+1}$$

$P(\sum_{f\in\mathcal{X}} B_f(t) \le U)$ *decrease exponentially*

- Finite horizon decision

$$U^{\pi^*}([t_i, t_{I+1}], \mathcal{X}_{i-1}) = \max_{\mathcal{X}_i} \{\overline{\Gamma}_s(i) - \Gamma_d(A_i) + U^{\pi^*}([t_{i+1}, t_{I+1}], \mathcal{X}_i)\}$$

Theorem: *Exact solution as a MILP*

Let $\Delta_i = t_{i+1} - t_i$. Every solution of the MILP

$$\max \sum_{i=1}^{I} \left\{ \Delta_i \left(\sum_{f\in\mathcal{F}} \overline{B}^i_f x_{i,f} - s_i \right) - \sum_{f\in\mathcal{F}} L_f b_{i,f} \right\}$$

s.t.

$$\sum_{f\in\mathcal{F}} \overline{B}^i_f x_{i,f} - s_i \le U, \quad \forall 1 \le i \le I \qquad (1)$$

$$x_{i,f} - x_{i-1,f} - b_{i,f} \le 0, \quad \forall 1 \le i \le I, f \in \mathcal{F} \qquad (2)$$

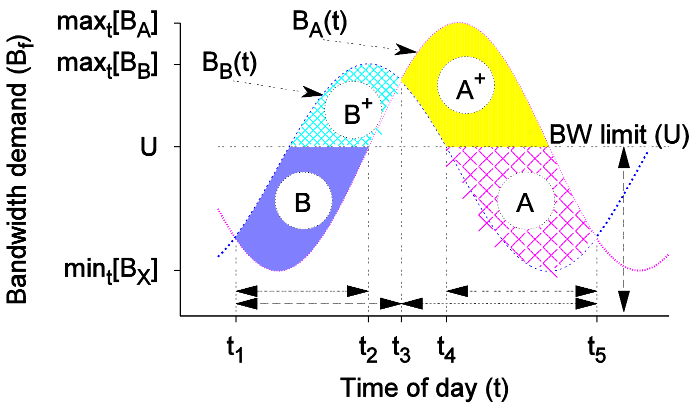
$$\sum_{f\in\mathcal{F}} L_f x_{i,f} \le S, \quad \forall 1 \le i \le I \qquad (3)$$

$$b_{i,f} \ge 0, \quad x_{i,f} \in \{0,1\}, \quad \forall 1 \le i \le I, f \in \mathcal{F} \qquad (4)$$

$$s_i \ge 0, \quad \forall 1 \le i \le I, \qquad (5)$$

is an optimal policy π^*.

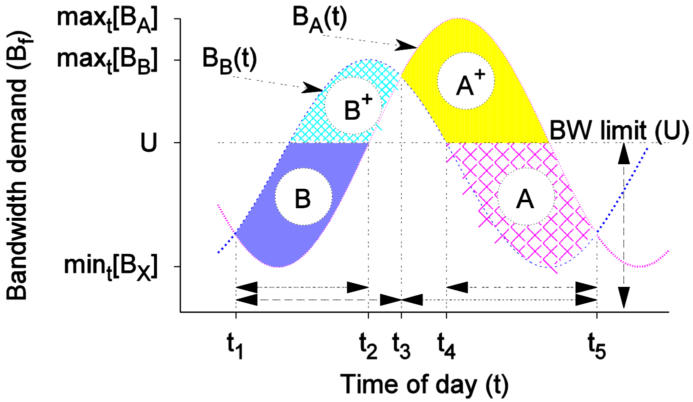
Policy: No Download Cost (NDC)



- Consider next interval only

$$\chi_i^{NDC} = \arg\max_{\chi_i} \overline{\Gamma}_s^{\pi}(i)$$

Policy: No Download Cost (NDC)



- Consider next interval only

$$\chi_i^{NDC} = \arg\max_{\chi_i} \overline{\Gamma}_s^{\pi}(i)$$

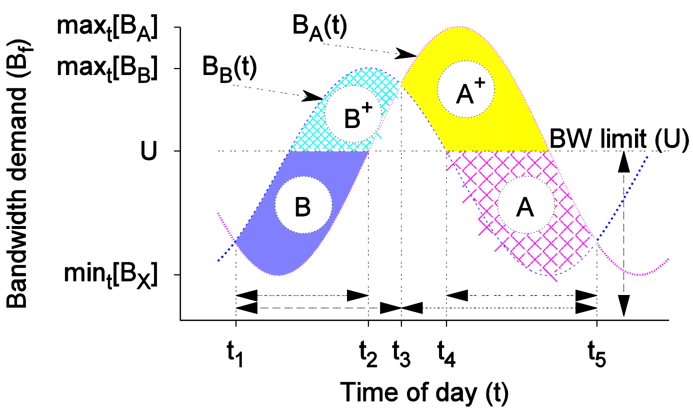
- Proposition 1: Unbounded approximation ratio

$$\frac{J^{NDC}}{J^{\pi^*}} = \frac{1+\epsilon}{1.5\epsilon} \Rightarrow \lim_{\epsilon\to 0} \frac{J^{NDC}}{J^{\pi^*}} = \infty$$

- Proposition 2: Approximation bound

The approximation ratio of NDC is $\frac{J^{NDC}}{J^{\pi^*}} \leq 1 + IS/J^{\pi^*}$.

Policy: No Download Cost (NDC)



- Consider next interval only

$$\chi_i^{NDC} = \arg\max_{\chi_i} \overline{\Gamma}_s^{\pi}(i)$$

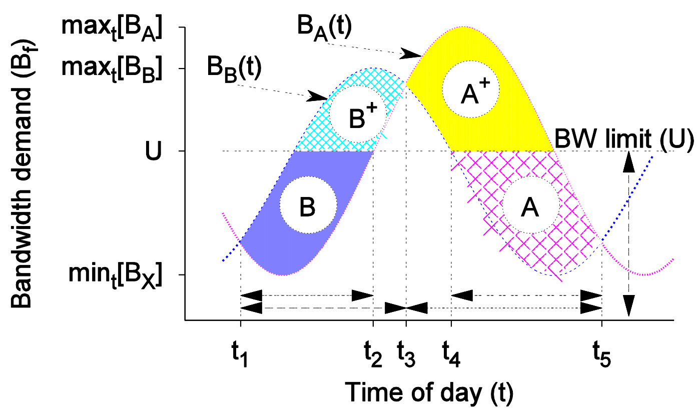
- Proposition 1: Unbounded approximation ratio

$$\frac{J^{NDC}}{J^{\pi^*}} = \frac{1+\epsilon}{1.5\epsilon} \Rightarrow \lim_{\epsilon \to 0} \frac{J^{NDC}}{J^{\pi^*}} = \infty$$

- Proposition 2: Approximation bound

The approximation ratio of NDC is $\frac{J^{NDC}}{J^{\pi^*}} \leq 1 + IS/J^{\pi^*}$.

Policy: No Download Cost (NDC)



- Consider next interval only

$$\chi_i^{NDC} = \arg\max_{\chi_i} \overline{\Gamma}_s^{\pi}(i)$$

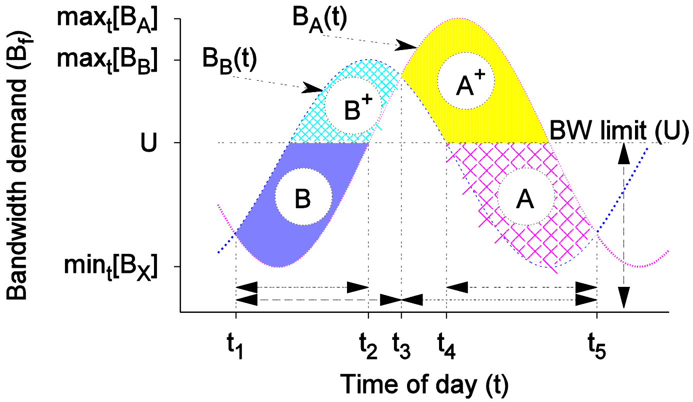
- Proposition 1: Unbounded approximation ratio

$$\frac{J^{NDC}}{J^{\pi^*}} = \frac{1+\epsilon}{1.5\epsilon} \Rightarrow \lim_{\epsilon\to 0} \frac{J^{NDC}}{J^{\pi^*}} = \infty$$

- Proposition 2: Approximation bound

The approximation ratio of NDC is $\frac{J^{NDC}}{J^{\pi^*}} \leq 1 + IS/J^{\pi^*}$.

Policy: No Download Cost (NDC)



- Consider next interval only

$$\chi_i^{NDC} = \arg\max_{\chi_i} \overline{\Gamma}_s^{\pi}(i)$$

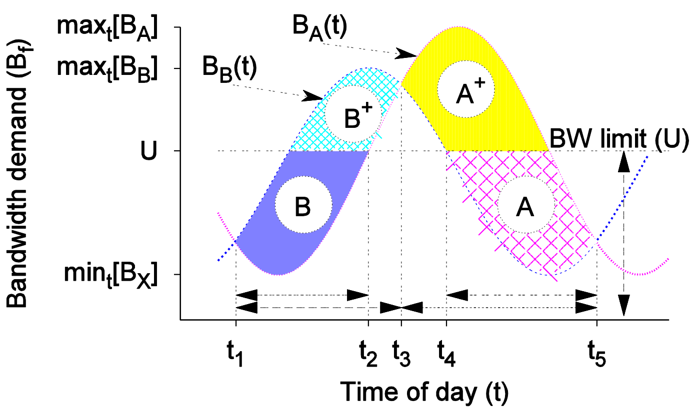
- Proposition 1: Unbounded approximation ratio

$$\frac{J^{NDC}}{J^{\pi^*}} = \frac{1+\epsilon}{1.5\epsilon} \Rightarrow \lim_{\epsilon \to 0} \frac{J^{NDC}}{J^{\pi^*}} = \infty$$

- Proposition 2: Approximation bound

The approximation ratio of NDC is $\frac{J^{NDC}}{J^{\pi^*}} \leq 1 + IS/J^{\pi^*}$.

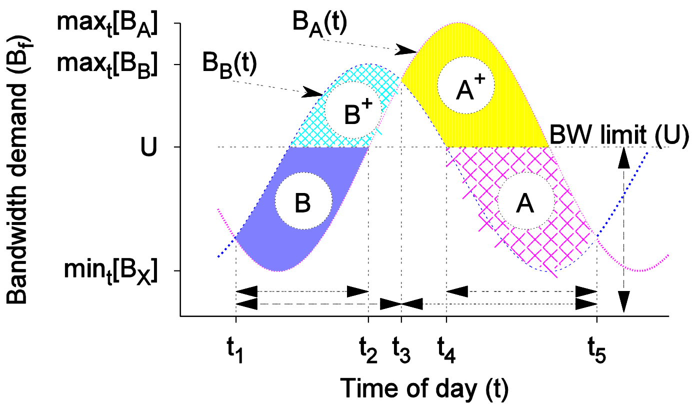
Policy: k-Step Look Ahead (k-SLA)



- Consider k next intervals

$$\chi_i^{1-SLA} = \arg\max_{\chi_i} \left\{ \overline{\Gamma}_s^{\pi}(i) - \Gamma_d^{\pi}(A_i) \right\}$$

Policy: k-Step Look Ahead (k-SLA)



- Consider k next intervals

$$\mathcal{X}_i^{1-SLA} = \arg\max_{\mathcal{X}_i} \left\{ \overline{\Gamma}_s^{\pi}(i) - \Gamma_d^{\pi}(A_i) \right\}$$

- Proposition 3: Unbounded approximation ratio

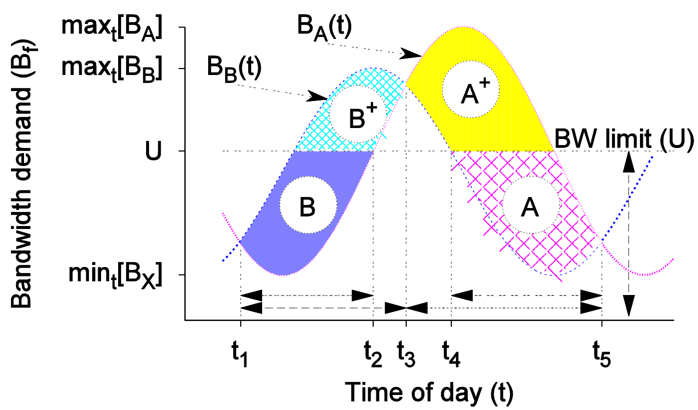
$$\frac{J^{1-SLA}}{J^{\pi^*}} = \frac{1+\epsilon}{3\epsilon} \Rightarrow \lim_{\epsilon \to 0} \frac{J^{1-SLA}}{J^{\pi^*}} = \infty$$

- Proposition 4: Approximation bound

Assume the average demand of each file inserted into the dedicated storage by an optimal policy π^* is lower bounded by a factor $\rho > 0$ such that the demand of each such file satisfies $\rho \frac{1}{I} \sum_{i=0}^{I} \overline{B}_f^i \Delta_i \geq L_f$. Then, for $k > \frac{\rho I}{I - \rho}$ the approximation ratio of k-SLA is

$$\frac{J^{k-SLA}}{J^{\pi^*}} \leq \frac{1}{1 - \frac{\rho}{k}(1 + \frac{k}{I})}. \qquad (6)$$

Policy: k-Step Look Ahead (k-SLA)



- Consider k next intervals

$$\chi_i^{1-SLA} = \arg\max_{\chi_i} \left\{ \overline{\Gamma}_s^{\pi}(i) - \Gamma_d^{\pi}(A_i) \right\}$$

- Proposition 3: Unbounded approximation ratio

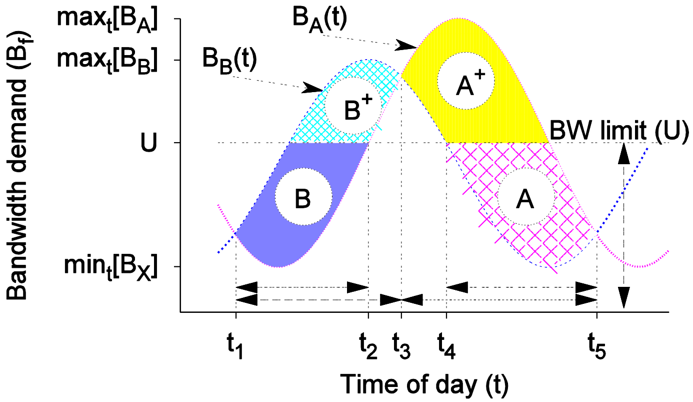
$$\frac{J^{1-SLA}}{J^{\pi^*}} = \frac{1+\epsilon}{3\epsilon} \Rightarrow \lim_{\epsilon \to 0} \frac{J^{1-SLA}}{J^{\pi^*}} = \infty$$

- Proposition 4: Approximation bound

Assume the average demand of each file inserted into the dedicated storage by an optimal policy π^* is lower bounded by a factor $\rho > 0$ such that the demand of each such file satisfies $\rho \frac{1}{I} \sum_{i=0}^{I} \overline{B}_f^i \Delta_i \geq L_f$. Then, for $k > \frac{\rho I}{I - \rho}$ the approximation ratio of k-SLA is

$$\frac{J^{k-SLA}}{J^{\pi^*}} \leq \frac{1}{1 - \frac{\rho}{k}(1 + \frac{k}{I})}. \qquad (6)$$

Policy: k-Step Look Ahead (k-SLA)



- Consider k next intervals

$$\chi_i^{1-SLA} = \arg\max_{\chi_i} \left\{ \overline{\Gamma}_s^{\pi}(i) - \Gamma_d^{\pi}(A_i) \right\}$$

- Proposition 3: Unbounded approximation ratio

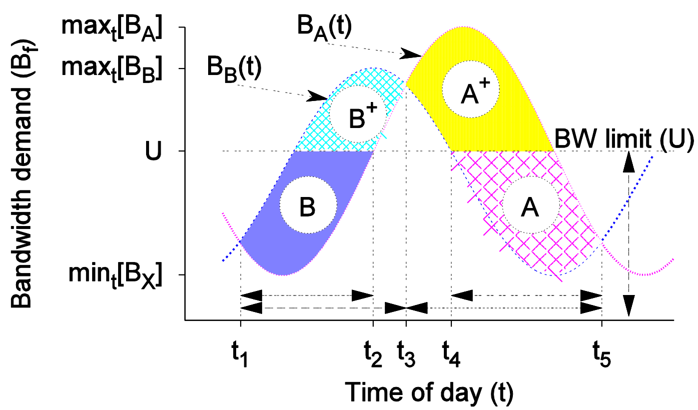
$$\frac{J^{1-SLA}}{J^{\pi^*}} = \frac{1+\epsilon}{3\epsilon} \Rightarrow \lim_{\epsilon \to 0} \frac{J^{1-SLA}}{J^{\pi^*}} = \infty$$

- Proposition 4: Approximation bound

Assume the average demand of each file inserted into the dedicated storage by an optimal policy π^* is lower bounded by a factor $\rho > 0$ such that the demand of each such file satisfies $\rho \frac{1}{I} \sum_{i=0}^{I} \overline{B}_f^i \Delta_i \geq L_f$. Then, for $k > \frac{\rho I}{I-\rho}$ the approximation ratio of k-SLA is

$$\frac{J^{k-SLA}}{J^{\pi^*}} \leq \frac{1}{1 - \frac{\rho}{k}(1 + \frac{k}{I})}. \qquad (6)$$

Policy: k-Step Look Ahead (k-SLA)



- Consider k next intervals

$$\chi_i^{1-SLA} = \arg\max_{\chi_i} \left\{ \overline{\Gamma}_s^{\pi}(i) - \Gamma_d^{\pi}(A_i) \right\}$$

- Proposition 3: Unbounded approximation ratio

$$\frac{J^{1-SLA}}{J^{\pi^*}} = \frac{1+\epsilon}{3\epsilon} \Rightarrow \lim_{\epsilon \to 0} \frac{J^{1-SLA}}{J^{\pi^*}} = \infty$$

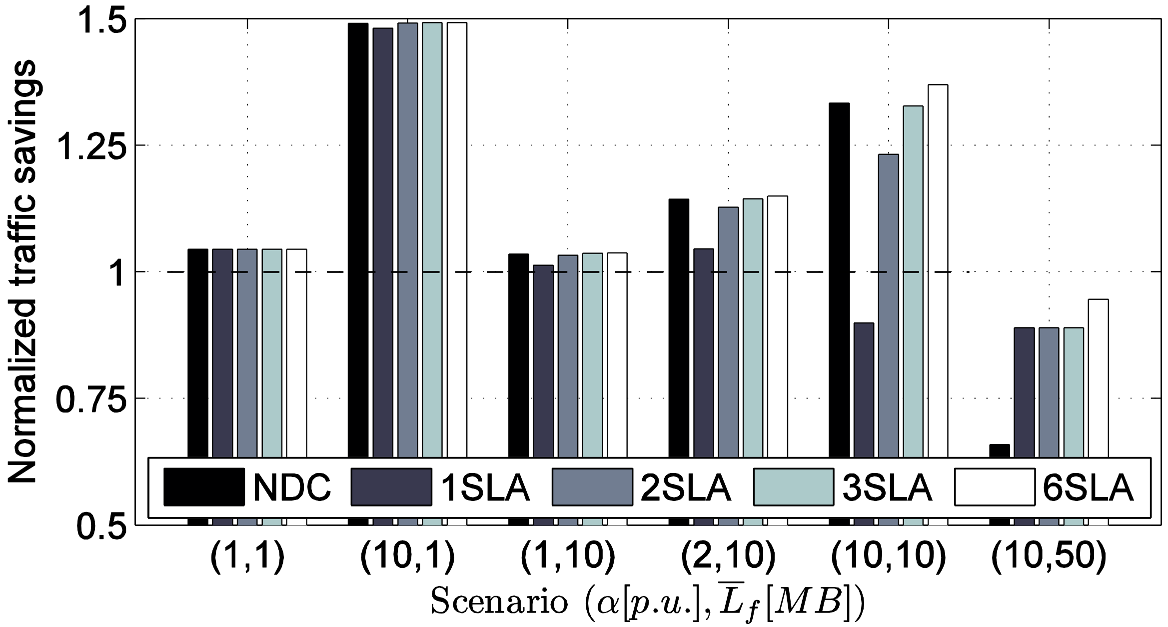
- Proposition 4: Approximation bound

Assume the average demand of each file inserted into the dedicated storage by an optimal policy π^* *is lower bounded by a factor* $\rho > 0$ *such that the demand of each such file satisfies* $\rho \frac{1}{I} \sum_{i=0}^{I} \overline{B}_f^i \Delta_i \geq L_f$. *Then, for* $k > \frac{\rho I}{I - \rho}$ *the approximation ratio of k-SLA is*

$$\frac{J^{k-SLA}}{J^{\pi^*}} \leq \frac{1}{1 - \frac{\rho}{k}(1 + \frac{k}{I})}. \qquad (6)$$

Trace-based analysis (Synthetic)

- Normalized traffic savings

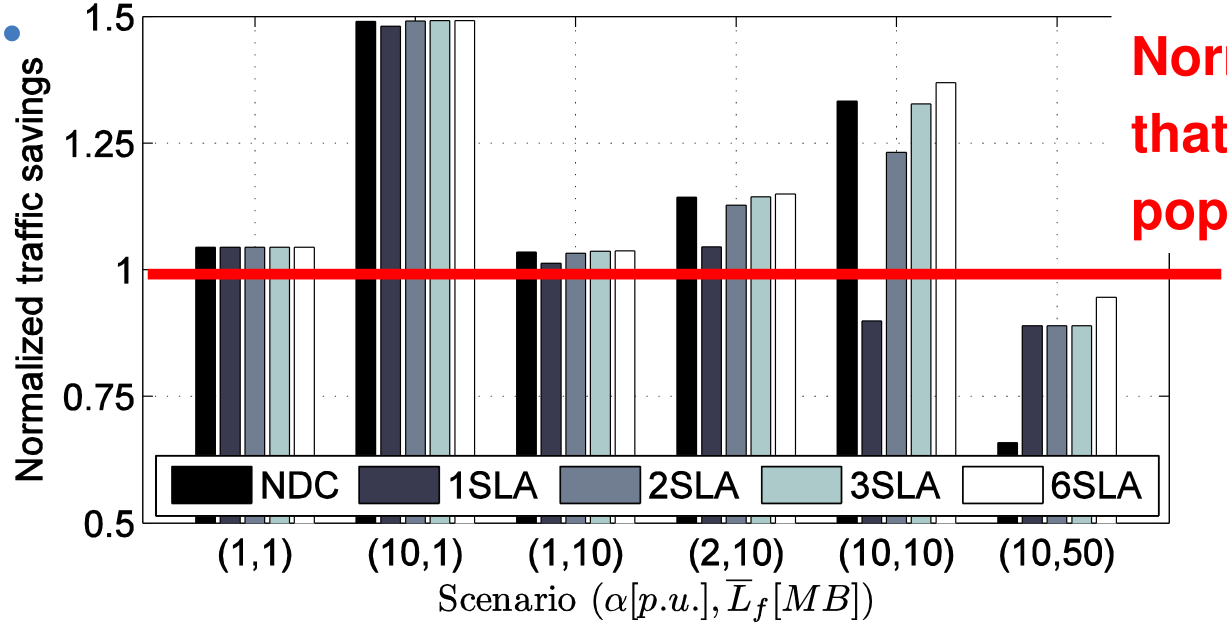


Based on Spotify trace characterization

- Workload: 3 groups of 1000 files; peaks N(0,2) offset by 8h for each group; sinusoid with 24h period; min/max ratio N(0.075,0.075), file sizes U(L/2,3L/2), bandwidth demand Bounded Pareto (B_{min}, B_{max}, α)

Trace-based analysis (Synthetic)

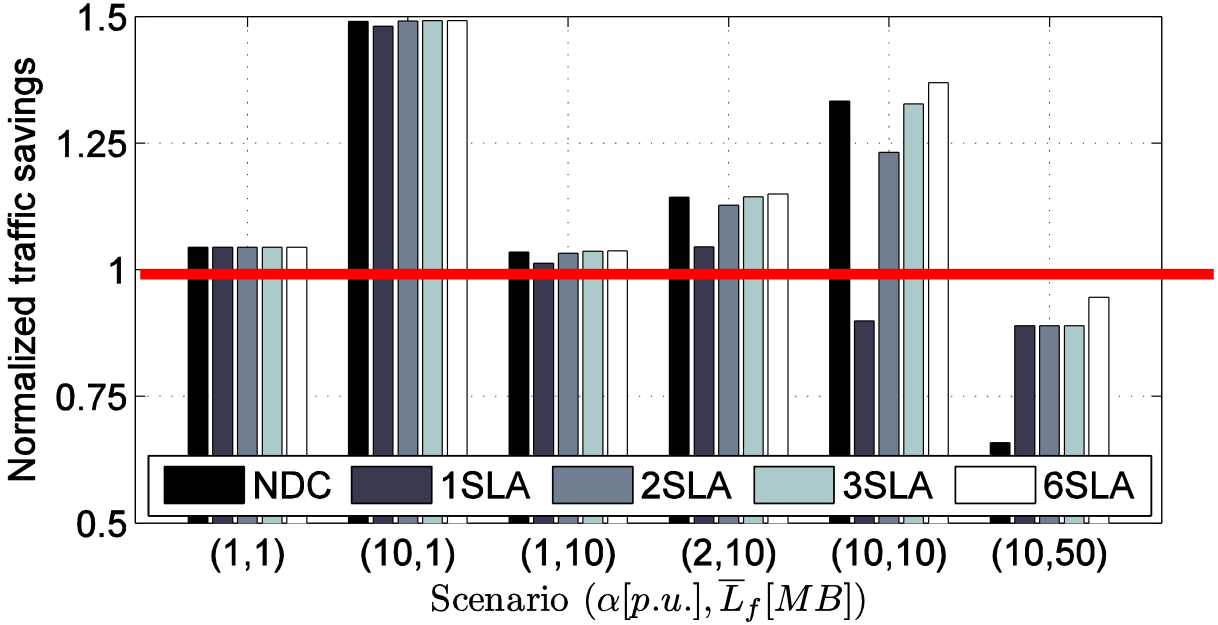
- Normalized traffic savings
-



Normalize against policy that stores the most popular files

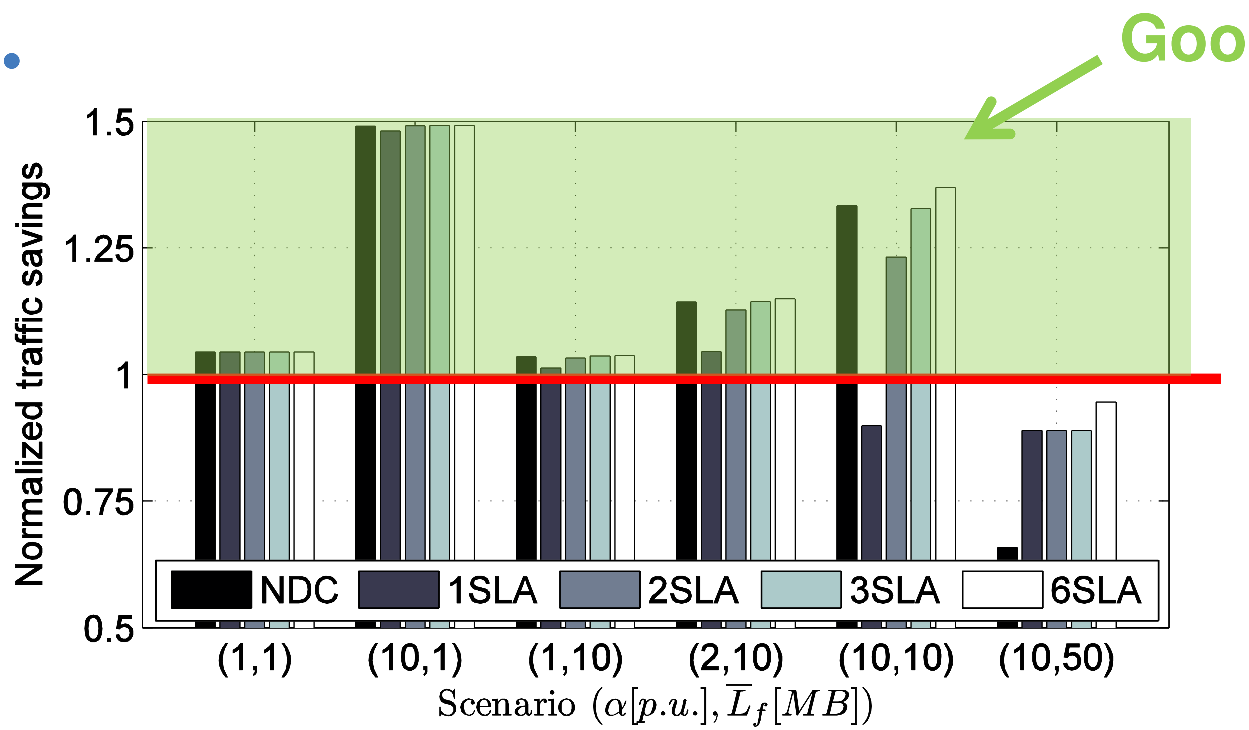
Trace-based analysis (Synthetic)

- Normalized traffic savings
-



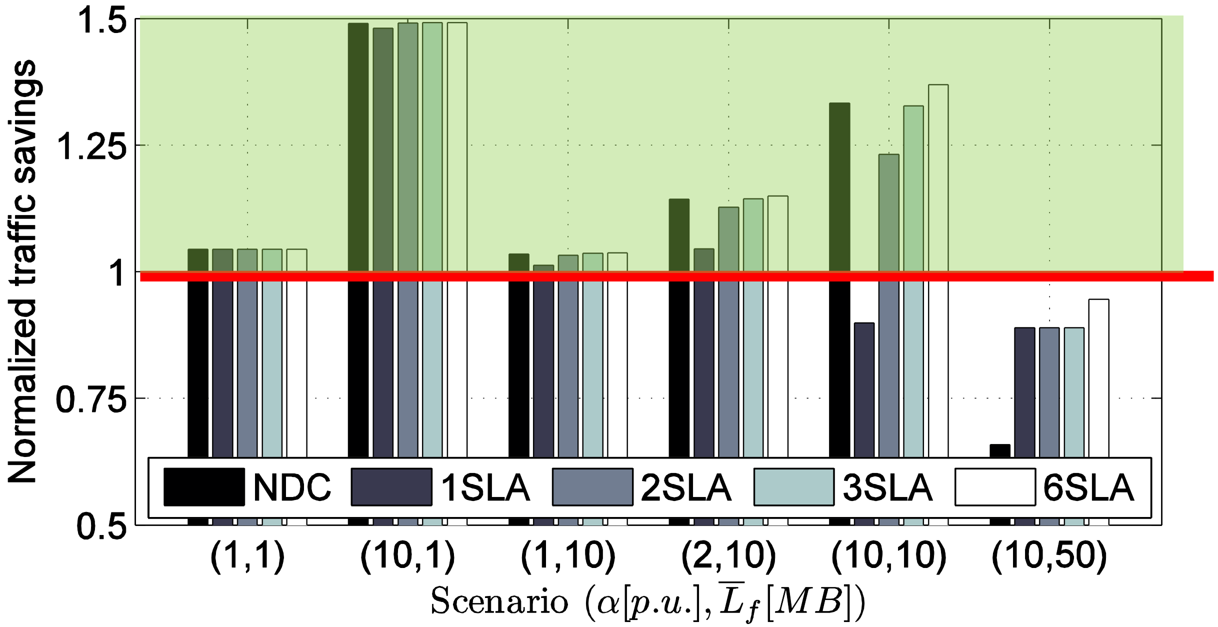
Trace-based analysis (Synthetic)

- Normalized traffic savings
-



Trace-based analysis (Synthetic)

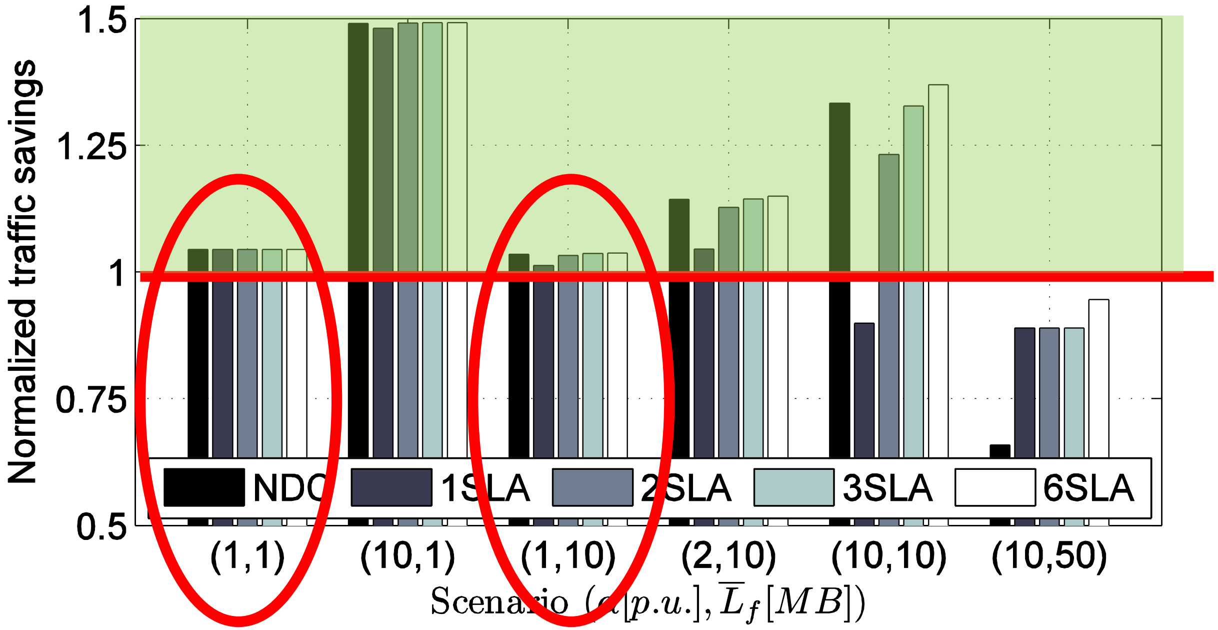
- Normalized traffic savings
-



- Modest gains when Zipf-like (α≈1) rank popularity
- Significant gains when more uniform (α≈10)
- NDC fails for larges sizes (6-SLA still works well)

Trace-based analysis (Synthetic)

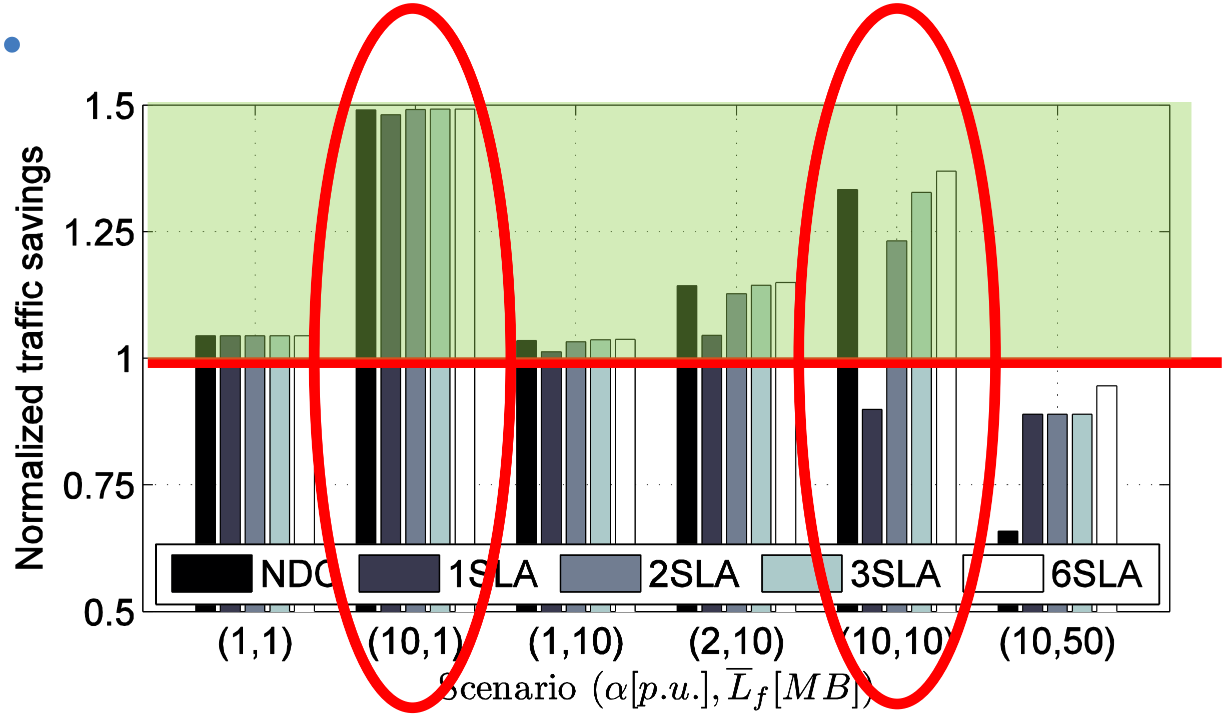
- Normalized traffic savings
-



- Modest gains when Zipf-like (α≈1) rank popularity
- Significant gains when more uniform (α≈10)
- NDC fails for larges sizes (6-SLA still works well)

Trace-based analysis (Synthetic)

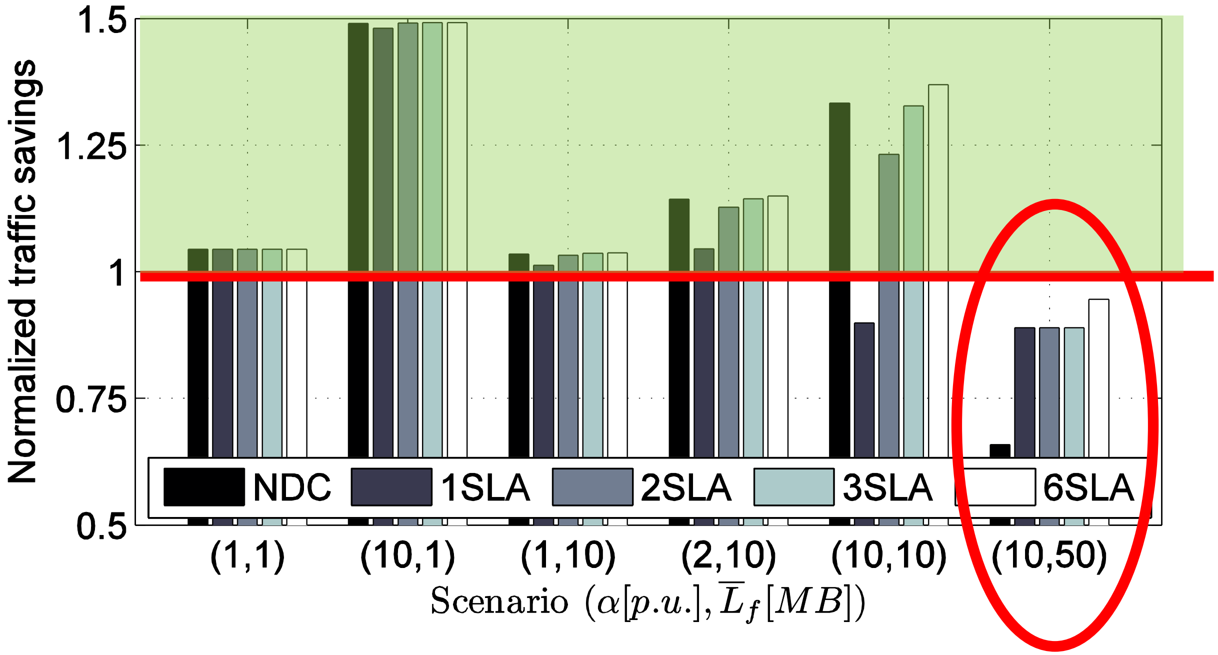
- Normalized traffic savings
-



- Modest gains when Zipf-like (α≈1) rank popularity
- Significant gains when more uniform (α≈10)
- NDC fails for larges sizes (6-SLA still works well)

Trace-based analysis (Synthetic)

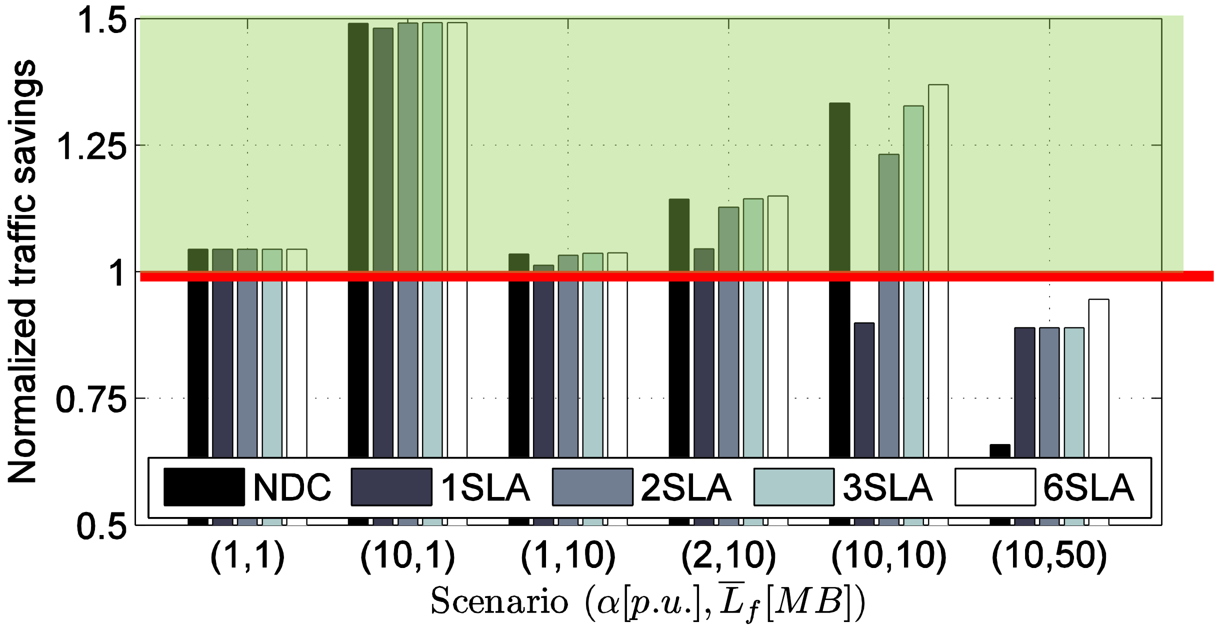
- Normalized traffic savings
-



- Modest gains when Zipf-like ($\alpha \approx 1$) rank popularity
- Significant gains when more uniform ($\alpha \approx 10$)
- NDC fails for larges sizes (6-SLA still works well)

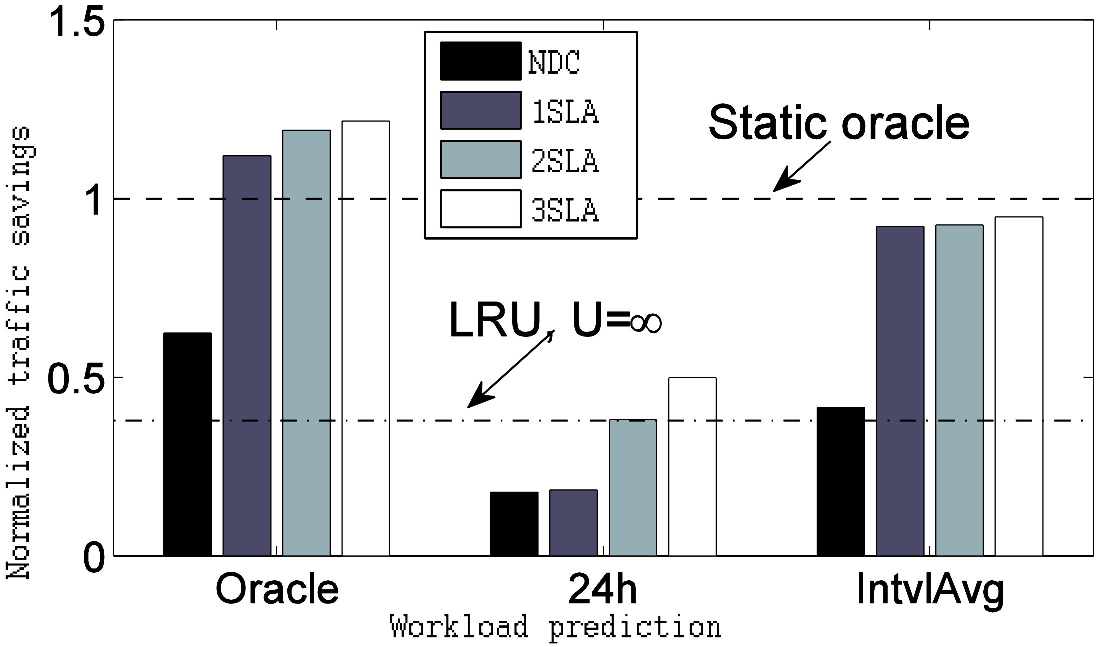
Trace-based analysis (Synthetic)

- Normalized traffic savings
-

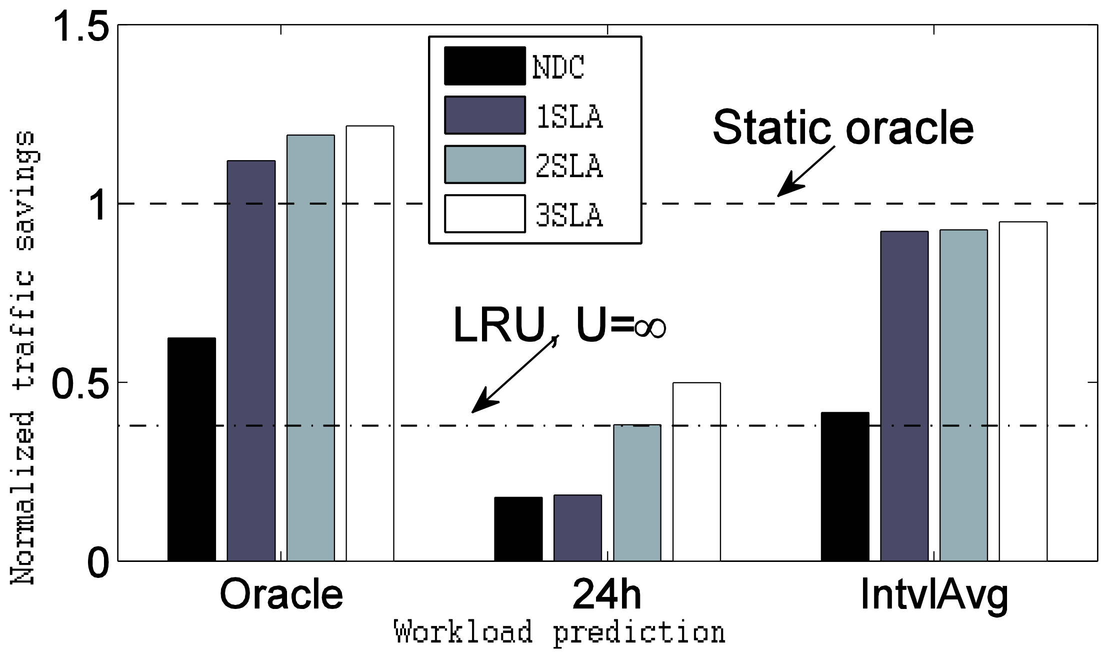


- Modest gains when Zipf-like ($\alpha\approx1$) rank popularity
- Significant gains when more uniform ($\alpha\approx10$)
- NDC fails for larges sizes (6-SLA still works well)

Trace-based Analysis

- Spotify traces (all requests for 1M random tracks; 1 week)
- Prediction policies: (i) "oracle", (ii) 24h, (iii) interval average
- 

Trace-based Analysis

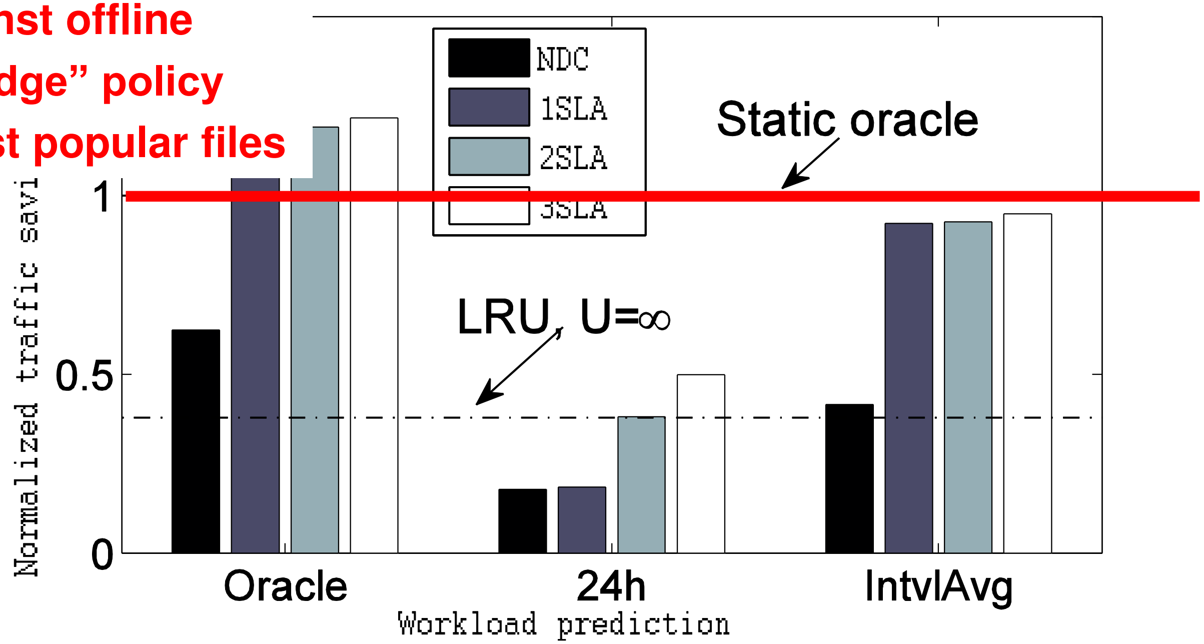
- Spotify traces (all requests for 1M random tracks; 1 week)
- Prediction policies: (i) “oracle”, (ii) 24h, (iii) interval average
- 

  - NDC fails; 3-SLA works reasonably well
  - Dynamic allocation with k-SLA outperform LRU by far

Trace-based Analysis

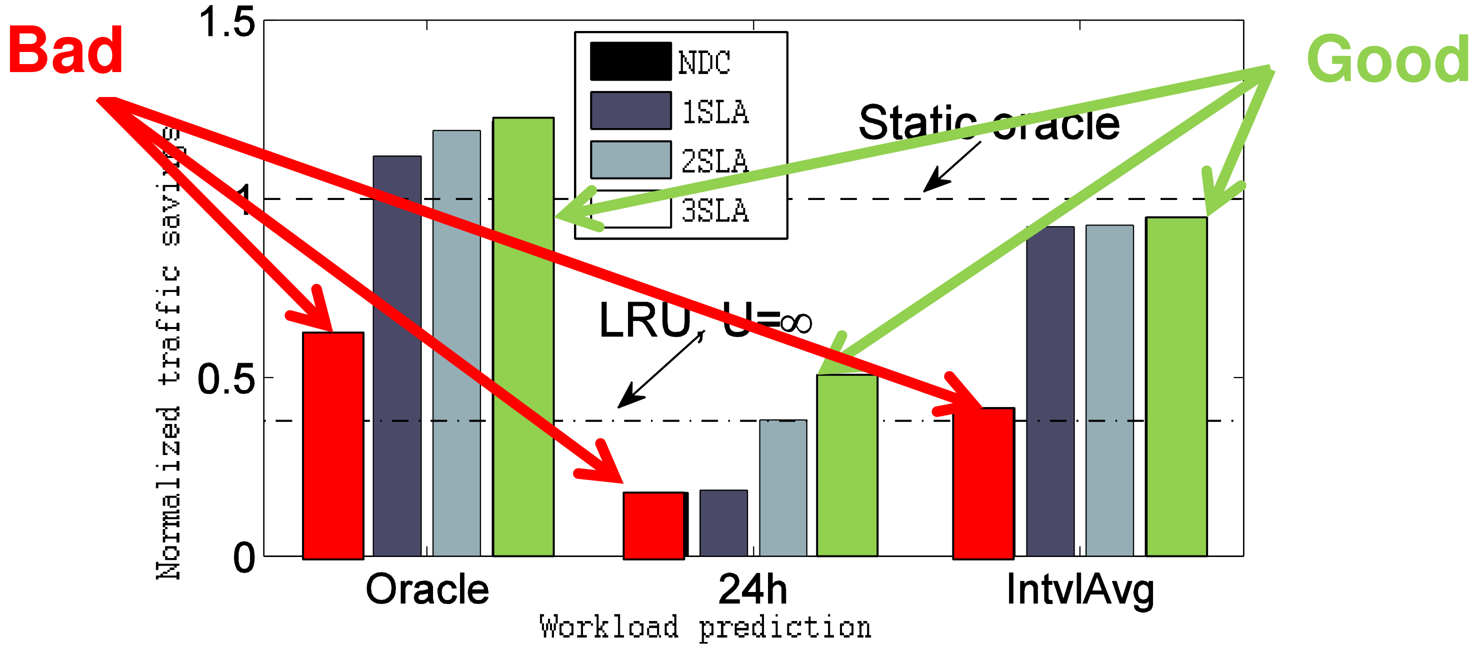
- Spotify traces (all requests for 1M random tracks; 1 week)
- Prediction policies: (i) “oracle”, (ii) 24h, (iii) interval average

Normalize against offline “global knowledge” policy that stores most popular files



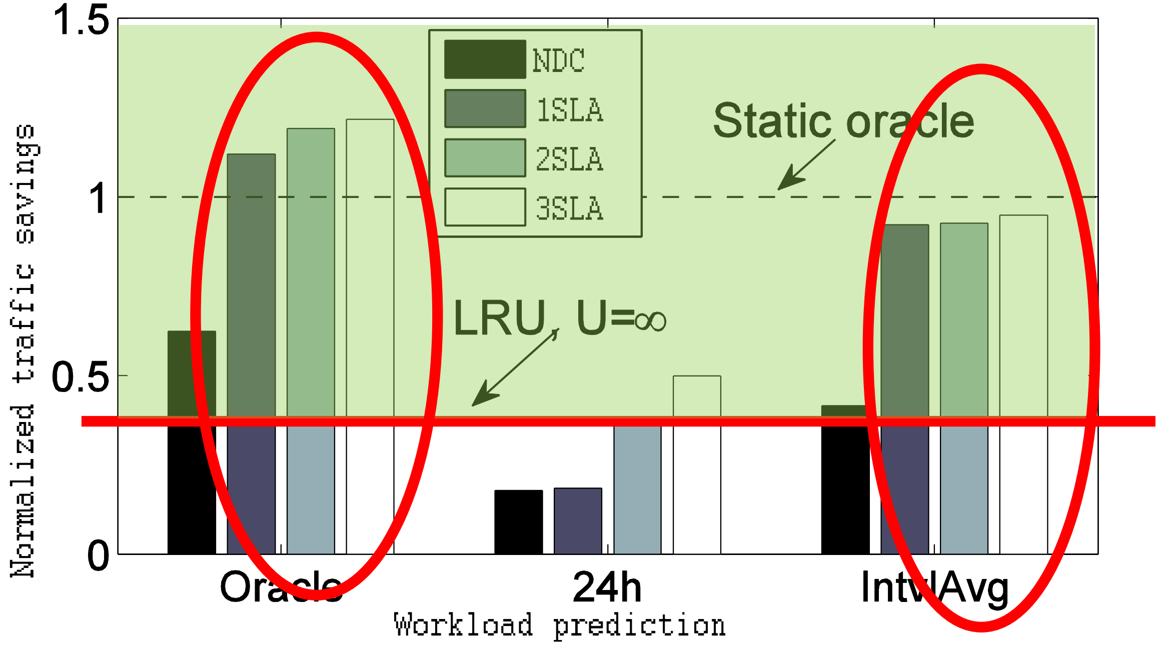
- NDC fails; 3-SLA works reasonably well
- Dynamic allocation with k-SLA outperform LRU by far

Trace-based Analysis

- Spotify traces (all requests for 1M random tracks; 1 week)
- Prediction policies: (i) “oracle”, (ii) 24h, (iii) interval average
- 

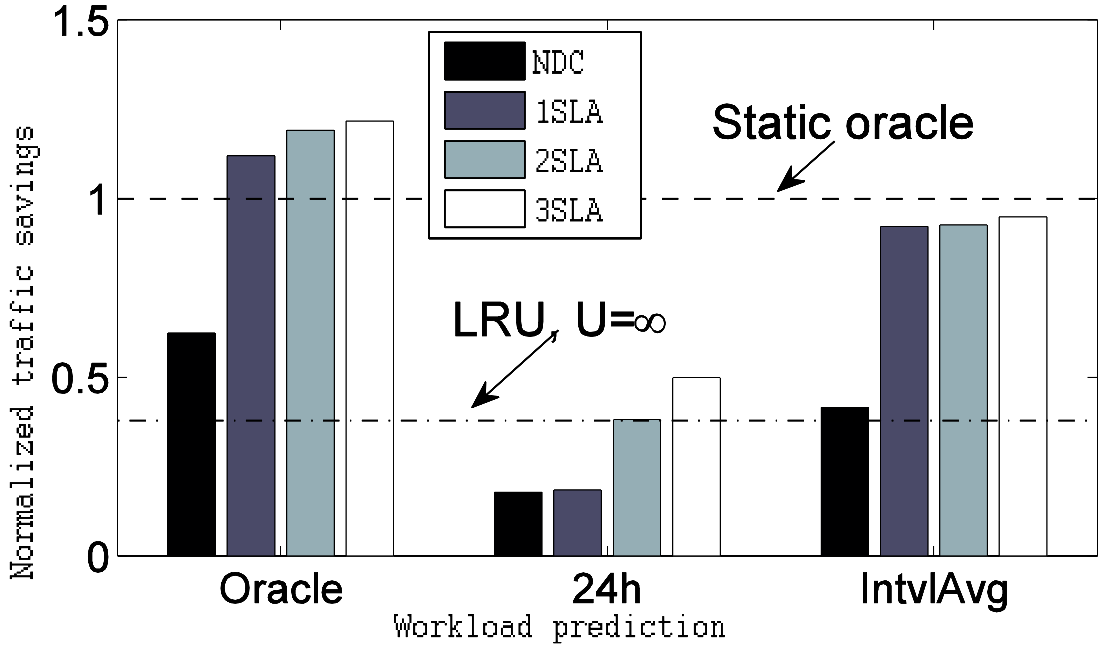
  - NDC fails; 3-SLA works reasonably well
  - Dynamic allocation with k-SLA outperform LRU by far

Trace-based Analysis

- Spotify traces (all requests for 1M random tracks; 1 week)
- Prediction policies: (i) “oracle”, (ii) 24h, (iii) interval average
- 

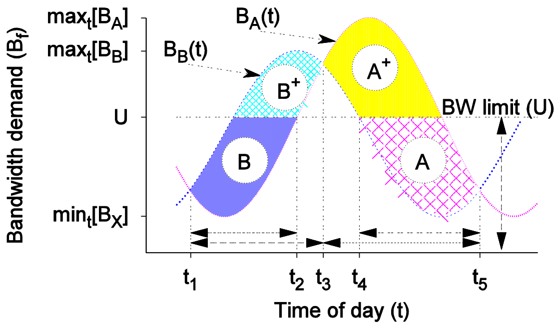
  - NDC fails; 3-SLA works reasonably well
  - Dynamic allocation with k-SLA outperform LRU by far

Trace-based Analysis

- Spotify traces (all requests for 1M random tracks; 1 week)
- Prediction policies: (i) “oracle”, (ii) 24h, (iii) interval average
- 

  - NDC fails; 3-SLA works reasonably well
  - Dynamic allocation with k-SLA outperform LRU by far

Dynamic Content Allocation Problem



- Finite horizon dynamic decision problem
- Discrete mean-value approximation
- Exact solution as MILP
- Computationally feasible approximations (e.g., k-SLA) with performance bounds
- Validate model and policies using traces from Spotify

LiU EXPANDING REALITY

Dynamic Content Allocation for Cloud-assisted Service of Periodic Workloads

György Dan (KTH) and Niklas Carlsson (LiU)

Thank you!

Niklas Carlsson (niklas.carlsson@liu.se)
www.ida.liu.se/~nikca/papers/infocom14.pdf