Understanding Service Demand for Adaptive Allocation of Distributed Resources

Jose Renato Santos, Yoshio Turner, John Janakiraman

HP Labs

Koustuv Dasgupta

University of Maryland at Baltimore County

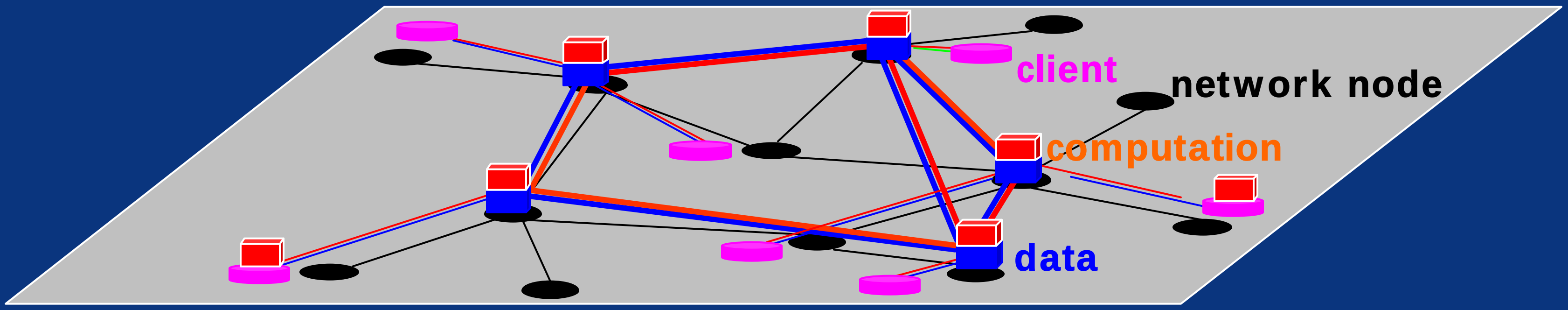
Outline

- Motivation
- Adaptive allocation of distributed resources
- Demand characterization
  - Demand distribution across clients
  - Regional distribution of clients and demand
- Conclusion

Motivation

- Internet Services characteristics:
  - High variation of demand (high peak/average ratio)
  - Demand is usually distributed over a wide area
  - High latency and low bandwidth over wide area
- Vision: Utility Computing model
  - Computing resources (servers, network bandwidth, storage) will be owned by infrastructure providers and dynamically allocated to service providers according to their current needs. (pay per use model)
  - Example: HP UDC (Utility Data Center) product

Adaptive Distributed Services



- Services will use distributed computing resources (wide area)
  - to reduce network latency to clients
  - to exploit resource markets
  - to harness distributed compute power
- Infrastructure needs to adapt dynamically
  - to satisfy service constraints
  - to respond to changes in demand and resource conditions

Adaptive allocation involves

- Selecting sites where services instances should be placed
- Controlling distribution of client demand to these service sites
- Allocating site resources proportionate to their demand
- Adapting these assignments as demand and resource conditions change

Factors influencing allocation decisions

- **Demand attributes**
  - **Location of clients**
  - **Demand intensity and distribution among clients**
- **Resource attributes**
  - **Available sites**
  - **Capacity (number of servers, storage, BW)**
  - **Cost**
- **Network attributes**
  - **Latency and BW from server sites to clients**
  - **Latency and BW among server sites**
- **Service attributes**
  - **Service requirements: Latency, disaster tolerance**
  - **Service characteristics: components, communication patterns among components, scalability properties, etc.**
- **Dynamic variations in these factors over time**

Demand characterization

- **Goal:**

Understand service demand characteristics important for resource allocation decisions

a) Understand how demand is distributed among clients

b) Understand how clients are distributed across the global Internet

Methodology

Data set:

- Web site for the 1998 Soccer World Cup
  - Duration: Web site active - 88 days, Event - 33 days
  - 1.3 billion hits
  - 2.7 million unique client IP addresses

Clustering:

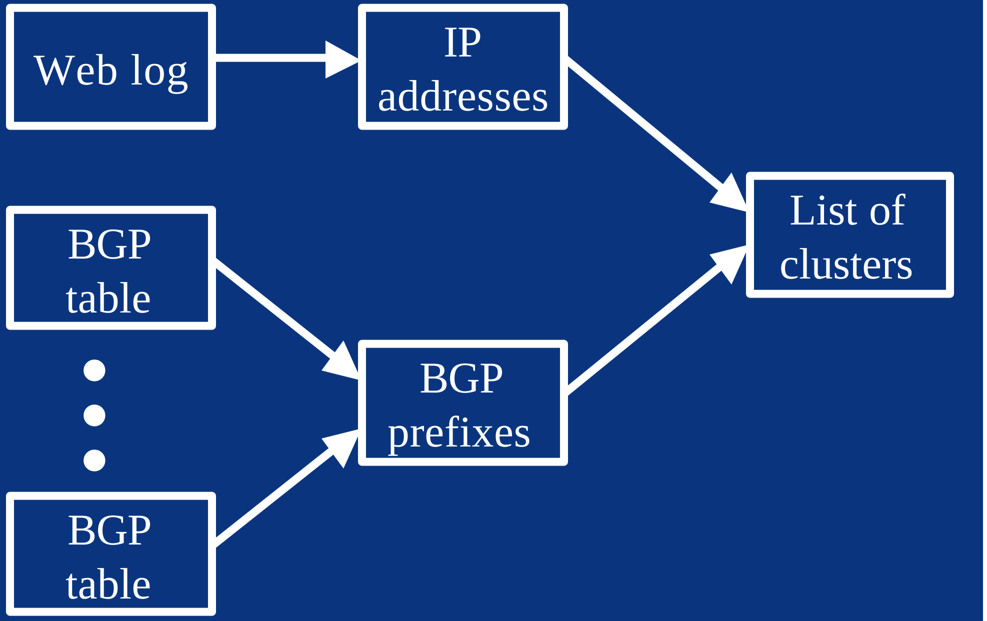
- Large number of clients
  - Difficult to analyze and interpret measurements
- Need to group clients in clusters
- Clustering should preserve topological distribution of clients
  - Clustering based on topological proximity

BGP client clustering

- **Technique proposed by Krishnamurthy & Wang [2000]**
  - **Based on BGP (Border Gateway Protocol) routing tables**
  - **Idea: clients that consistently share BGP routes are close to each other**

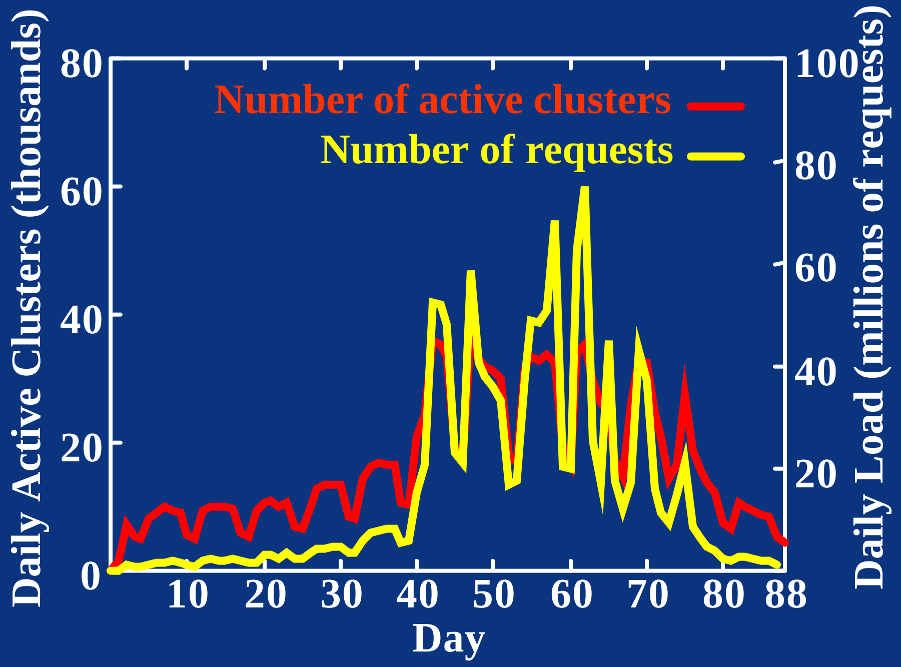
BGP table (route across AS's)

IP prefix/mask	Next hop



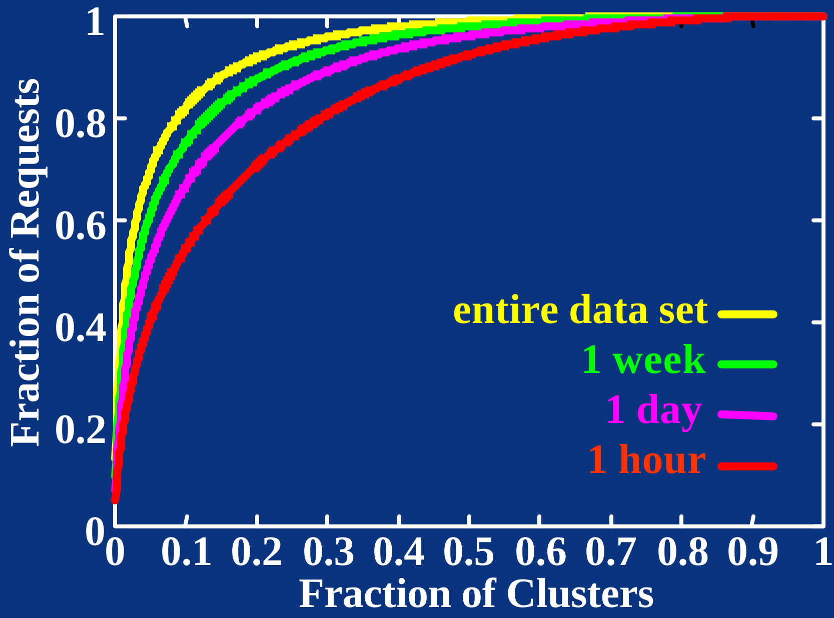
- **Result: 2.7 million clients → 81,420 clusters**

Daily demand variation



- **demand varies significantly over time**
  - **dynamic allocation of resources is beneficial**

Demand variation among clusters



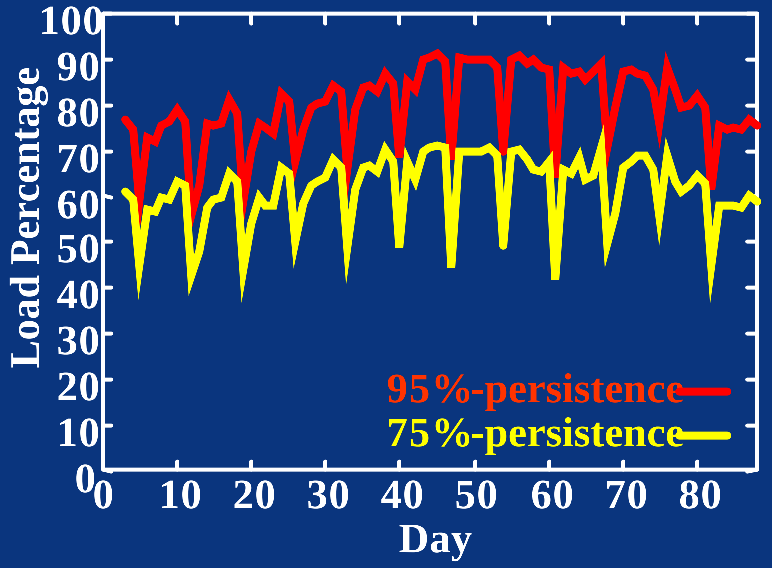
- 20% of clusters contribute 90% of overall World Cup requests

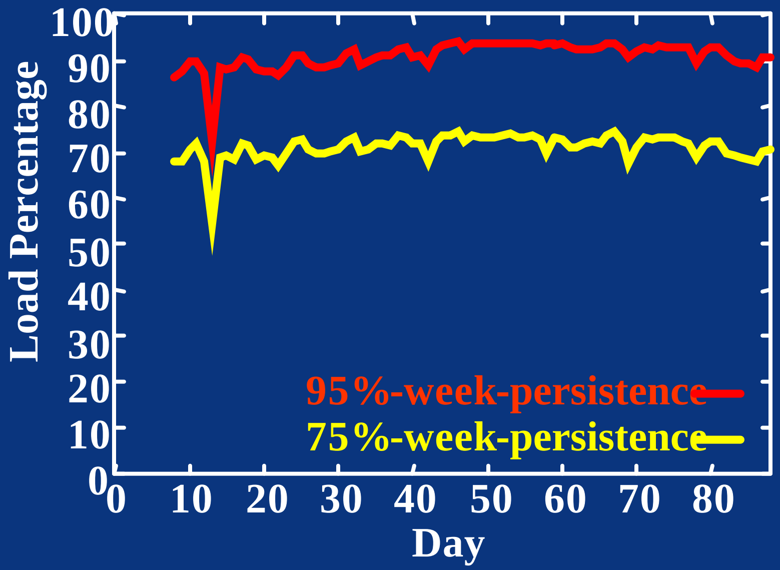
- **Skewed load: A few clusters contribute to majority of load**

  ⇨**monitoring/probing only a small subset of clusters is sufficient to characterize demand**

Predictability of dominant set of clusters

- ***p%-persistent clusters:*** intersection of set of most active clusters generating p% of load on a given day with the similar set for the previous day
- ***p%-week-persistent clusters:*** intersection of set of most active clusters generating p% of load on a given day with the similar sets for the previous 7 days





- **Active clusters are predictable from recent history**
  - ⇨ **useful for good placement**

Demand characterization

- **Goal:**

Understand Service Demand characteristics important for resource allocation decisions

a) **Understand how demand is distributed among clients**

b) **Understand how clients are distributed across the global Internet (Regional demand)**

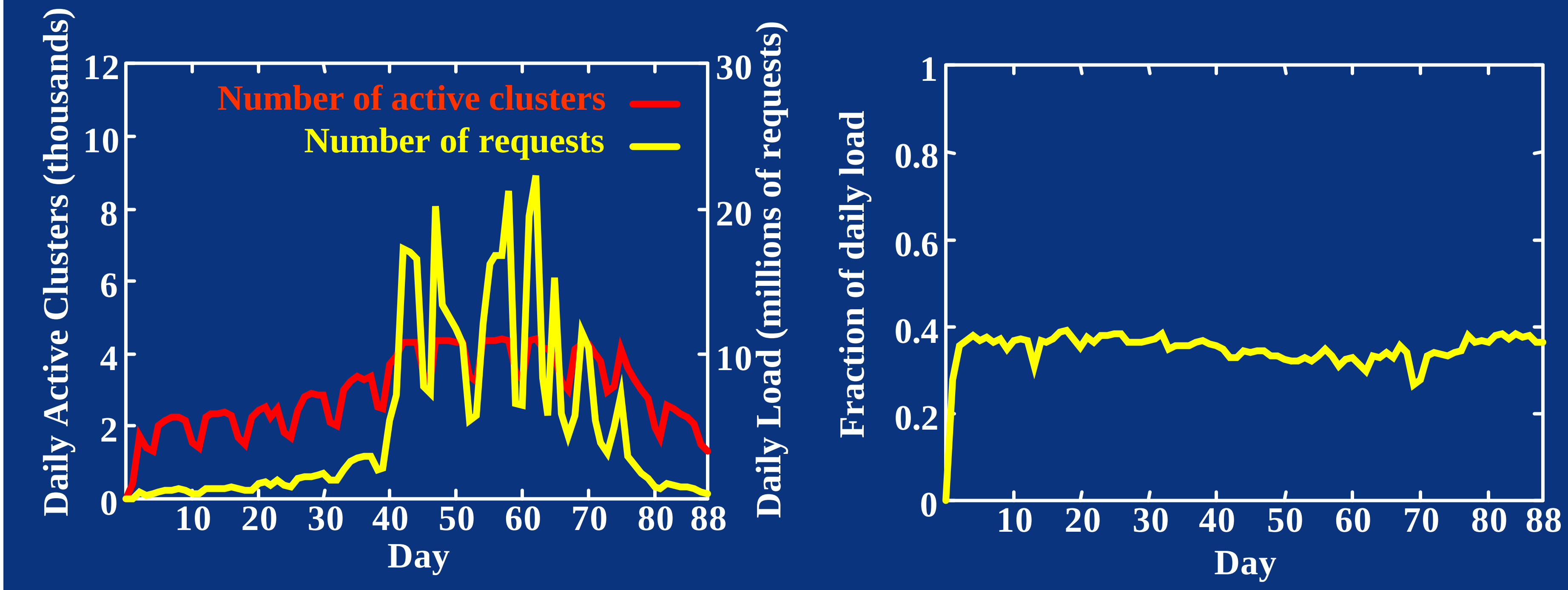
Regional demand methodology

Subdivide global internet in large regions

- Used 17 ping servers distributed around the world for defining 17 regions
  - North America: 7, Europe: 8, Africa: 1, Australia: 1
- Selected Subset of clusters
  - Dominant clusters responsible for 90% of load
- Group clusters in 17 non-overlapping regions
  - Estimated cluster/server latency using “ping”
  - Assign each cluster to the region of “closest” ping server

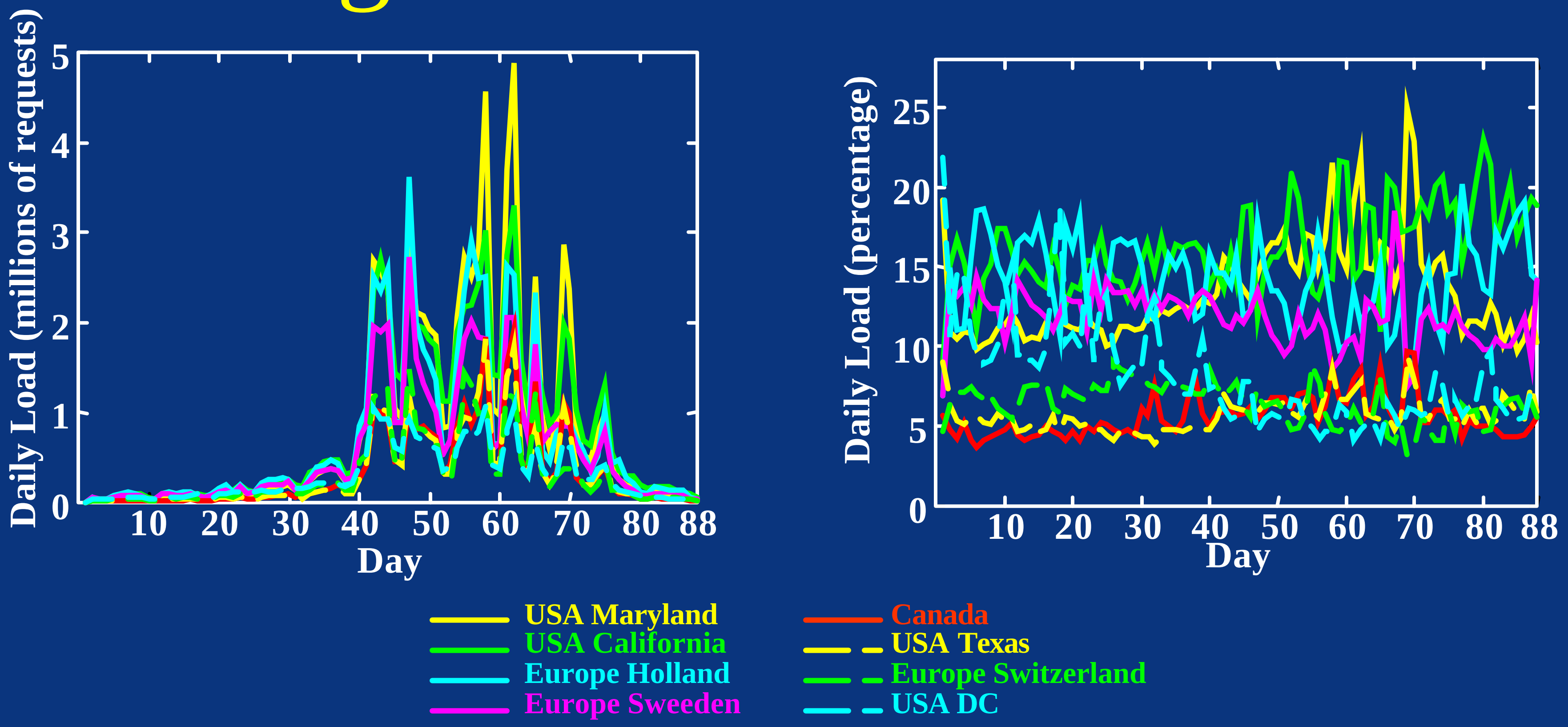
Clusters used in regional demand study

- **Only a subset of clusters was consistently reachable in the experiments**



- **Load pattern of subset is a scaled version of original**
  - **40% of original**

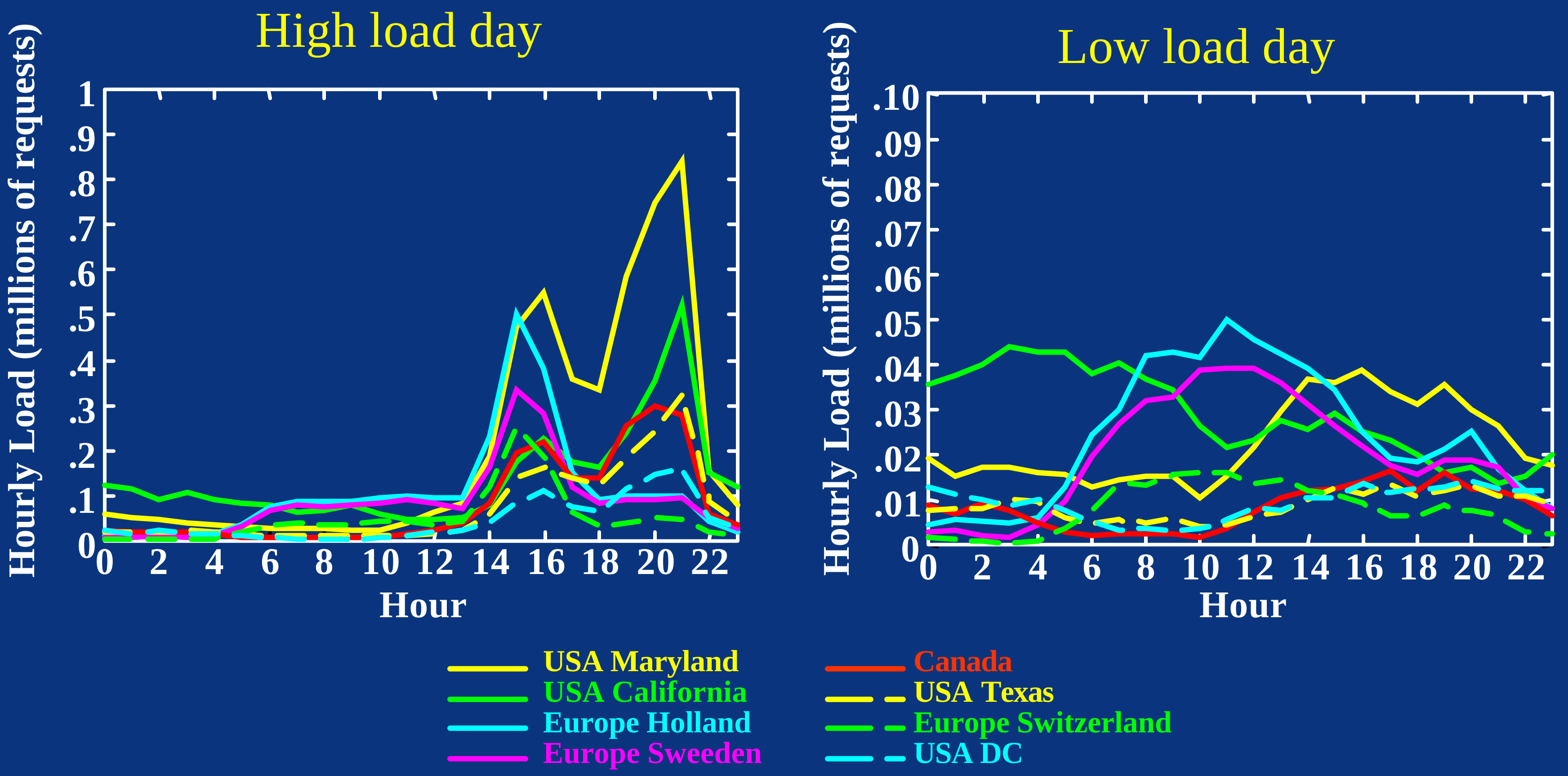
Regional load distribution



- Small changes in relative load despite large changes in absolute load

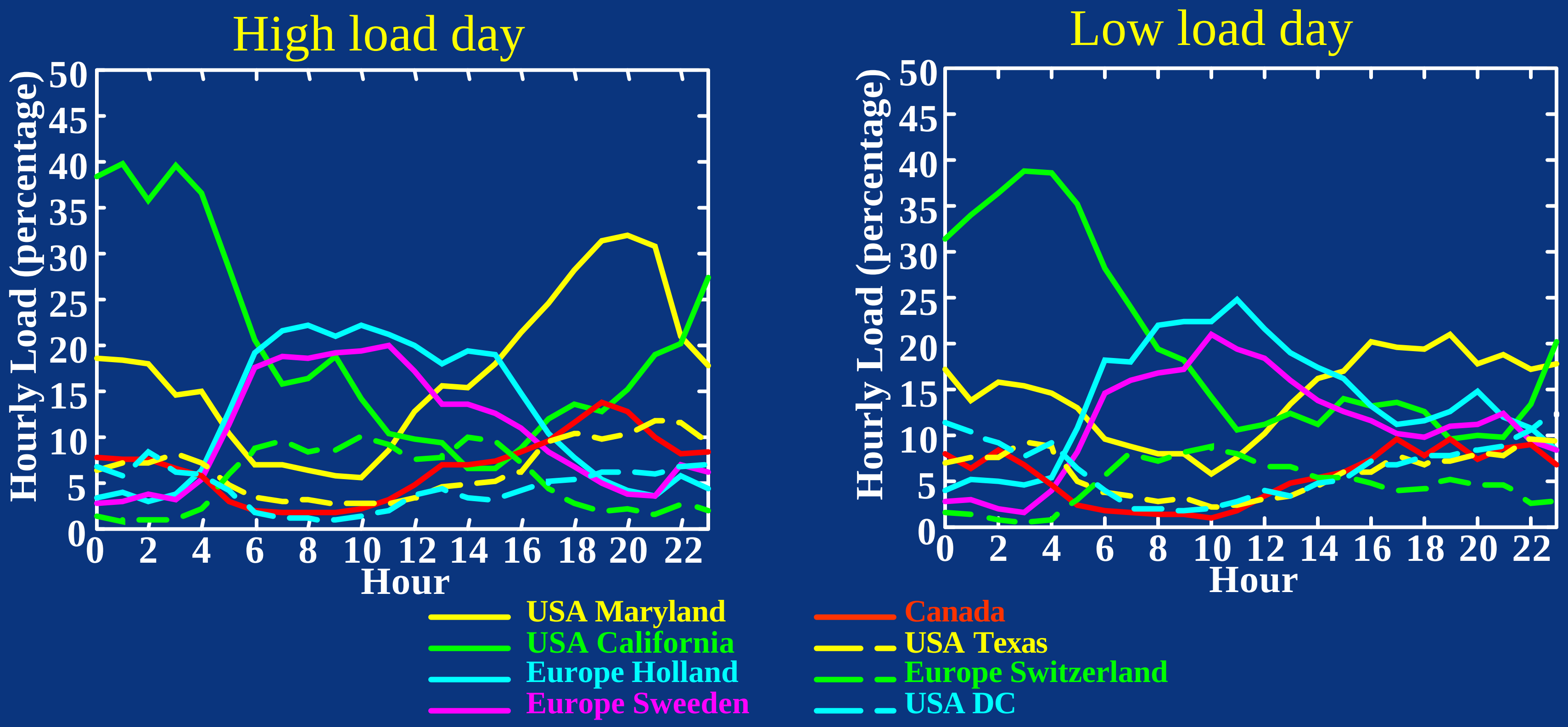
  ⇨ Regional distribution of load predictable (even if total load is not)

Regional demand - hourly distribution



- **Different absolute load patterns**

Regional demand - hourly distribution



- **Relative load of regions varies from hour to hour in any day**
  - ⇨ **dynamic placement/routing may be beneficial**
- **Similar pattern of hourly variations on multiple days (time zone)**
  - ⇨ **dynamics of hourly pattern can be predicted**

Conclusion

- Studied demand characteristics of the 1998 World Cup Web site (for service placement)
- Small subset of clusters dominates demand
  - Stable on a daily basis (Useful for good placement)
- Dynamic allocation is desirable
  - Particularly to scale up/down resource allocation at each site
  - Dynamic changes in resource placement may be beneficial in some cases (To handle hourly demand variations)
    - Variations are predictable (Resources could be reserved)
- Need to consider other factors (service requirements, resource costs, resource characteristics variations, etc.) to make allocation decisions
- Other workloads may have different characteristics

hp
®
invent