On Demand Systems

ISR Annual Research Review

Dr. Alfred Z. Spector

Vice President, Services & Software Research

IBM Corporation

aspector@us.ibm.com

June 17, 2003

IBM®

ISR Annual Research Review

Abstract

Over the 50 years of modern computer science, computer systems have had a demonstrated capacity to automate an enormous variety of tasks, and per-tasks costs have been greatly reduced. However, there are a two key challenges on the horizon: 1. In many areas, further declines in transaction costs by traditional means are subject to the laws of diminishing returns. 2. The complexity of infrastructure management threatens to outweigh the benefits of further automation. In this talk, I shall illustrate these two dilemmas and describe a research agenda aimed at them. One foundation of this agenda is process integration with a heavy focus on *continual optimization* -- the application of mathematical techniques to optimize operations at many systemic level and at varying granularities of time. The other foundation is *autonomic computing* -- worked aimed at automating automation. I shall survey some research projects at IBM that are related to these two areas, but attempt also to more broadly describe the overall territory.

aspector@us.ibm.com

IBM Research Division

aspector@us.ibm.com

6/17/03

Outline

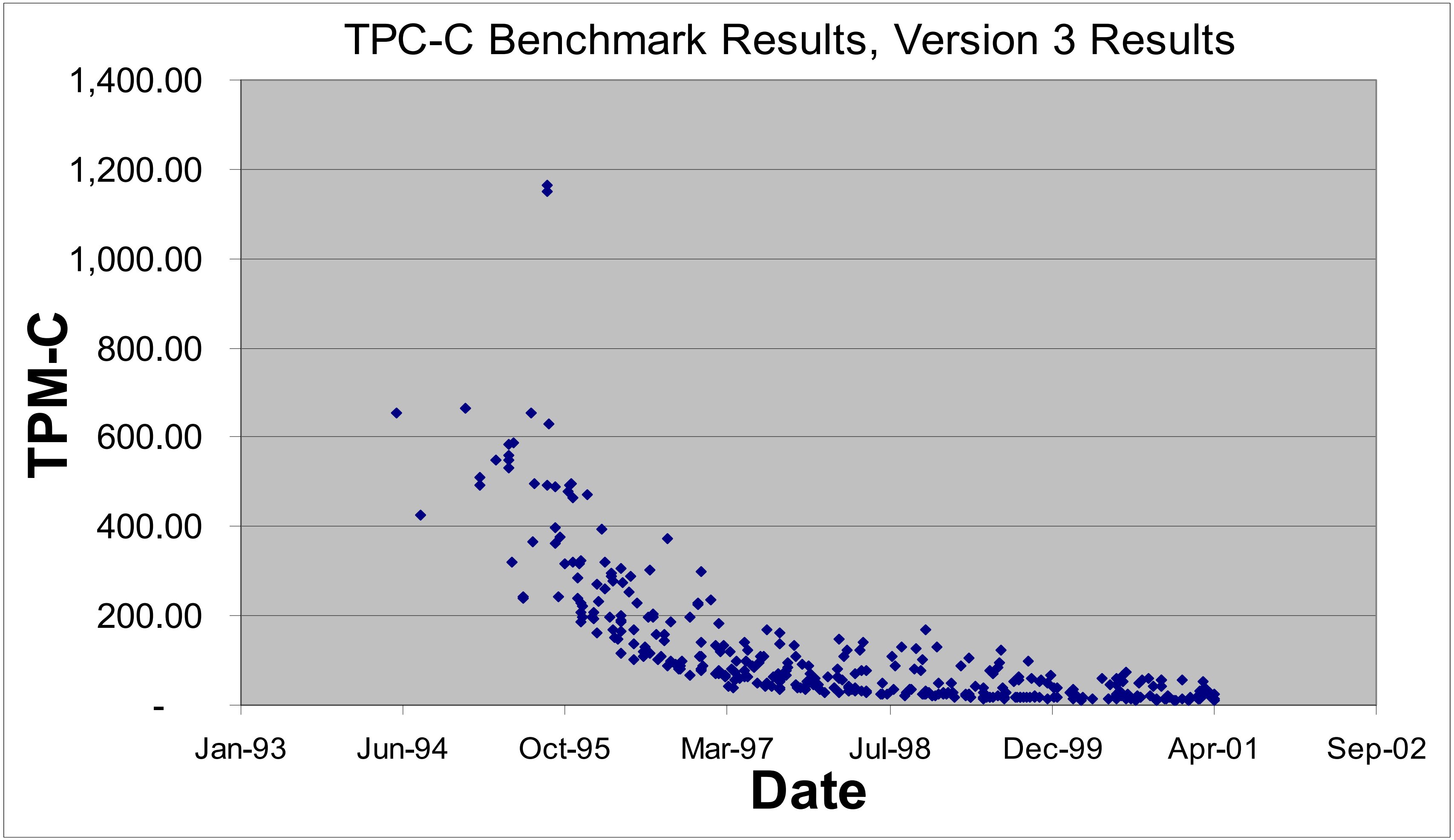
- Computer Science and I/T: Enormous Success, But Two Problems Lurk
  - Complexity
  - Diminishing Returns in Traditional Computing
- Hence, A Portion of IBM's On Demand Research Agenda:
  - Autonomic Computing
  - Continual Optimization

aspector@us.ibm.com

I/T Success

- Incalculable Benefits
- A key fact has been enormous decline in per-transaction costs
  - From Saber to the Web
  - Increasing
    - Flexibility & Function
  - Decreasing
    - Cost
- But, there are limits to these benefits

From WWW.TPC.ORG



aspector@us.ibm.com

ISR Annual Research Review

However, Complexity is Rising

- This seems intuitively right, but
  - What do I mean?
  - What evidence is there?

aspector@us.ibm.com

ISR Annual Research Review

3 Categories of Complexity

- Classic Complexity
  - Time
  - Space
- Implementation Complexity
  - Logical
  - Structural
  - Comprehensibility
- Usage Complexity

Task	Pre-Use	Novice	Middle	Expert	Exception
Install					
Configure					
Administer					
Use					

aspector@us.ibm.com

Usage Complexity Must Be Focus

- Consider a humble table (that is, legs and a horizontal surface)
  - *Classic Complexity* is not relevant as defined but there may be parallels
  - *Implementation Complexity* very high
    - Physicists do not fully understand tables, I suspect
  - *Usage Complexity* very low
- While Classic & Implementation Complexity may impact Usage Complexity, they are less important end goals. (In effect, they are tools of Computer Science.)

Software Systems Today

- Score high on most metrics:
  - Amount of code
  - # of dependencies
  - # of programmatic interfaces
  - # of layers
  - Administrative interface size & configuration options
  - Non-uniformity
  - Non-orthogonality
  - Defects
  - Documentation
  - # of programmers involved

Anecdotes on Systems

- Implementation Complexity:
  - Windows XP Code is tens of millions of lines of code supposedly with circa $10^5 – 10^6$ bugs
  - Cisco Routers have support for more than 100 protocols
- Usage (Administrative) Complexity: Sendmail
  - Access.db, domaintable.db, local-host-names, mailertable.db, submit.cf, …
  - More than a page of Features, defines, etc for sendmail.cf
  - Longest O'Reilly book, at ~1200 pages
- Usage (Use) Complexity: New BMW 7 Series
  - "But why did those Bavarian motor masters have to ruin a wonderful driving machine by filling it with more gadgets and gizmos than you'd find in the cockpit of a space shuttle? The only thing intuitive is how to open the doors," *Milwaukee Journal Sentinel*, 4/19/2002.

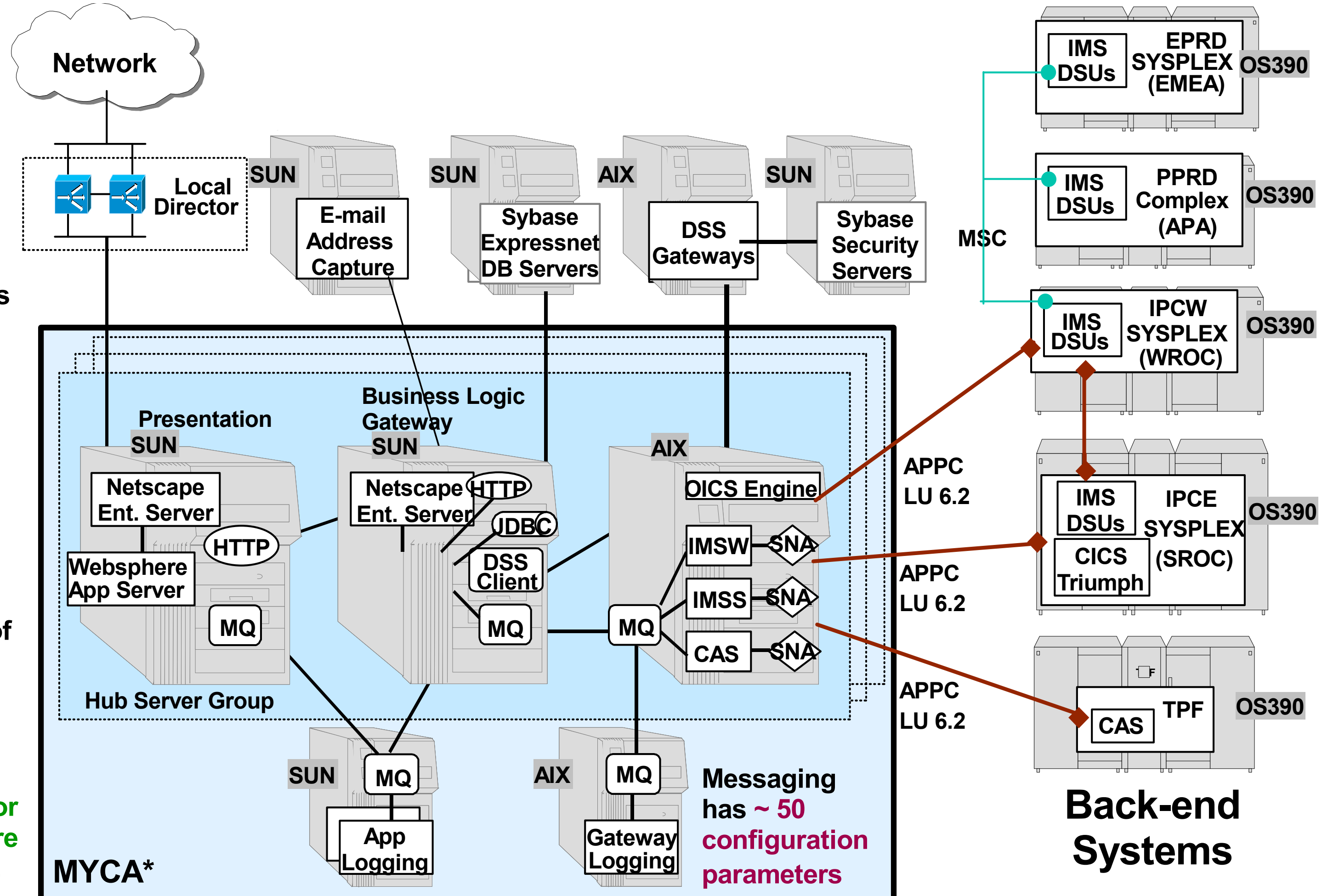
ISR Annual Research Review

Example: Credit Card Processing

An application server typically supports
- 5 Applications
- 10 EJBs
- Hundreds of servlets
- ~ 100 configuration parameters

A web server typically serves
- Thousands of web artifacts
- ~ 20 configuration parameters

Failure protocols for each component are different: time-out, number of retries, where and what they log, how they fail



* my credit card account

aspector@us.ibm.com

6/17/03

IBM

ISR Annual Research Review

RosettaNet Purchase Orders

- There are 551 XML fields in the PurchaseOrderRequest
- There are 700 XML fields in the PurchaseOrderConfirmation

Excerpted First lines of purchase order confirmation:

```
fromRole.PartnerRoleDescription
     |-- ContactInformation
     |    |-- contactName.FreeFormText
     |    |-- EmailAddress
     |-- facsimileNumber.CommunicationsNumber
     |    |-- telephoneNumber.CommunicationsNumber
     |-- GlobalPartnerRoleClassificationCode
     |-- PartnerDescription
     |    |-- BusinessDescription
     |    |    |-- GlobalBusinessIdentifier
     |    |    |-- GlobalSupplyChainCode
     |    |-- GlobalPartnerClassificationCode
GlobalDocumentFunctionCode
PurchaseOrder
     |-- AccountDescription
     |    |-- accountName.FreeFormText
     |    |-- AccountNumber
     |    |-- billTo.PartnerDescription
```

```
|    |    |-- BusinessDescription
|    |    |    |-- businessName.FreeFormText
|    |    |    |-- GlobalBusinessIdentifier
|    |    |    |-- PartnerBusinessIdentification
|    |    |    |    |-- ProprietaryBusinessIdentifier
|    |    |    |    |-- ProprietaryDomainIdentifier
|    |    |    |    |-- ProprietaryIdentifierAuthority
|    |    |-- ContactInformation
|    |    |    |-- contactName.FreeFormText
|    |    |    |-- EmailAddress
|    |    |    |-- facsimileNumber.CommunicationsNumber
|    |    |    |-- PhysicalLocation
|    |    |    |    |-- GlobalLocationIdentifier
|    |    |    |    |-- PartnerLocationIdentification
|    |    |    |    |    |-- ProprietaryDomainIdentifier
|    |    |    |    |    |-- ProprietaryIdentifierAuthority
...
```

Note: RosettaNet is a consortium of major companies working to create and implement industry-wide, open e-business process standards, that will form a common e-business language, globally aligning processes between supply chain partners. (From RosettaNet Home Page.)

Partial motivation: e-business on demand

- Dramatically decrease administrative complexity of information technology
  - One approach: automation of automation
    - Hence, *Autonomic computing*
- Dramatically increase value of information technology
  - Focus less on the technology
  - More on the impact of technology on the world
    - Hence, *Continual Optimization*

Autonomic Computing

aspector@us.ibm.com

6/17/03

IBM®

Autonomic Computing Vision

- "Intelligent" open systems that…
  - Manage complexity
  - "Know" themselves
  - Continuously tune themselves
  - Adapt to unpredictable conditions
  - Prevent & recover from failures
  - Provide a safe environment

Frees businesses to focus on business, not infrastructure

Autonomic Computing Benefits

- Increased return on IT investment (ROI)
  - Lower administrative costs
  - Higher asset utilization
  - IT alignment with business goals
  - Increased performance
- Improved resiliency: Quality of Service (QoS)
  - Reduced downtime
  - Better security
- Faster implementation of new capabilities: Time to Value (TTV)
  - Faster / more accurate installation
  - Fewer test cycles

Autonomic Self-Management

Increase Responsiveness

Adapt to dynamically changing environments

Business Resiliency

Discover, diagnose, act to prevent disruptions

Operational Efficiency

Tune resources, balance workloads to best use IT resources

Secure Information & Resources

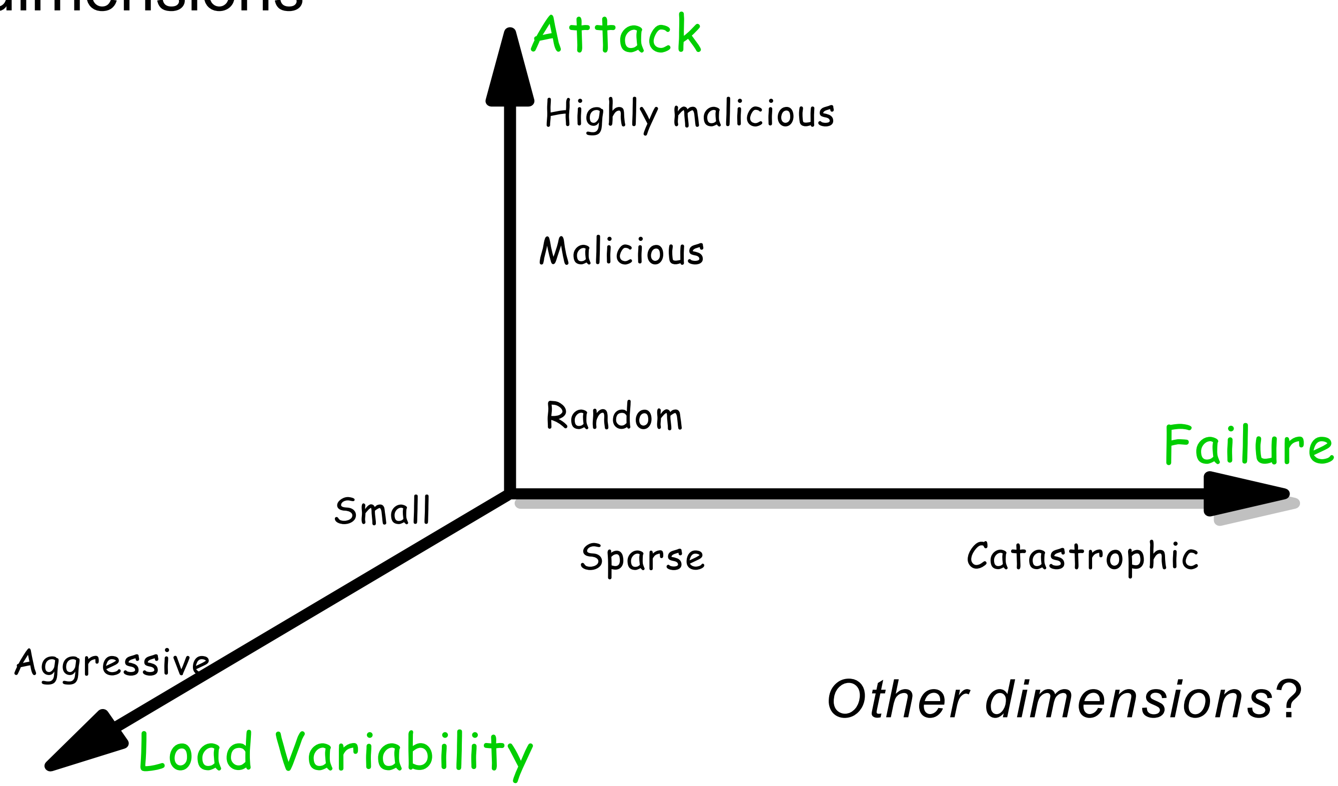
Anticipate, detect, identify, deter attacks

aspector@us.ibm.com

IBM®

Science and Technology

- Autonomic computing concept: Making systems robust in the presence of stimuli occurring in different dimensions



Other dimensions?

aspector@us.ibm.com

IBM

Evolving to Autonomic Computing

	Basic Level 1	**Managed** Level 2	**Predictive** Level 3	**Adaptive** Level 4	**Autonomic** Level 5
Characteristics	Multiple sources of system generated data				
Skills	*Extensive, highly skilled* IT staff				
Benefits	Basic Requirements Met				

Manual → **Autonomic**

aspector@us.ibm.com

ISR Annual Research Review

Evolving to Autonomic Computing

	Basic Level 1	Managed Level 2	Predictive Level 3	Adaptive Level 4	Autonomic Level 5
Characteristics	Multiple sources of system generated data	Data & actions consolidated through mgt tools			
Skills	*Extensive, highly skilled* IT staff	IT staff *analyzes & takes actions*			
Benefits	Basic Requirements Met	Greater system awareness Improved productivity			

Manual → Autonomic

aspector@us.ibm.com

6/17/03

Evolving to Autonomic Computing

	Basic Level 1	Managed Level 2	Predictive Level 3	Adaptive Level 4	Autonomic Level 5
Characteristics	Multiple sources of system generated data	Data & actions consolidated through mgt tools	Sys monitors correlates & recommends actions		
Skills	*Extensive, highly skilled* IT staff	IT staff *analyzes & takes actions*	IT staff *approves & initiates actions*		
Benefits	Basic Requirements Met	Greater system awareness Improved productivity	Less need for deep skills Faster/better decision making		

Manual → Autonomic

aspector@us.ibm.com

IBM®

Evolving to Autonomic Computing

	Basic Level 1	Managed Level 2	Predictive Level 3	Adaptive Level 4	Autonomic Level 5
Characteristics	Multiple sources of system generated data	Data & actions consolidated through mgt tools	Sys monitors, correlates & recommends actions	Sys monitors, correlates & takes action	
Skills	*Extensive, highly skilled* IT staff	IT staff *analyzes & takes actions*	IT staff *approves & initiates actions*	IT staff *manages performance* against SLAs	
Benefits	Basic Requirements Met	Greater system awareness Improved productivity	Less need for deep skills Faster/better decision making	Human/system interaction IT agility & resiliency	

Manual → Autonomic

aspector@us.ibm.com

Evolving to Autonomic Computing

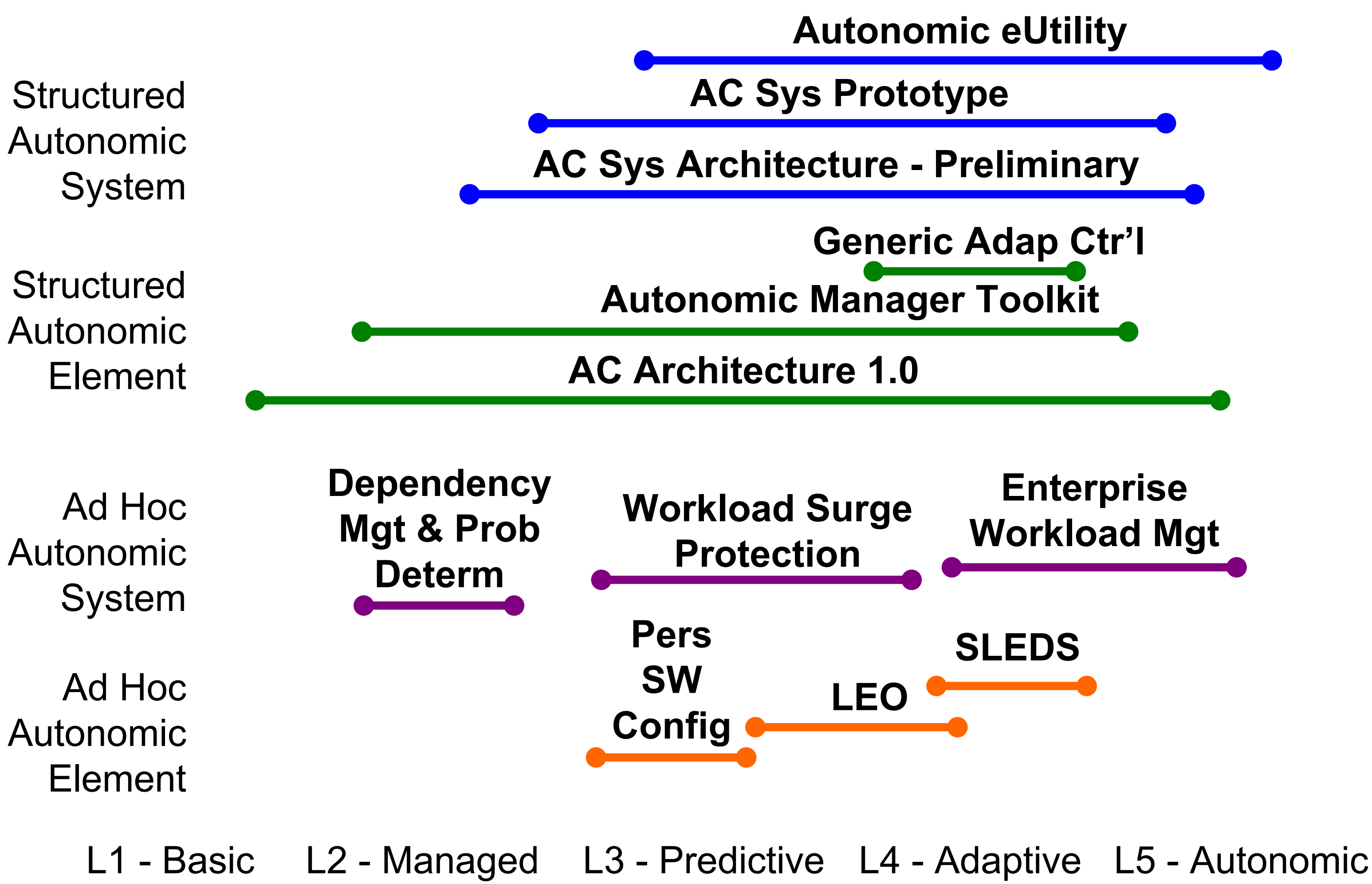
	Basic Level 1	Managed Level 2	Predictive Level 3	Adaptive Level 4	Autonomic Level 5
Characteristics	Multiple sources of system generated data	Data & actions consolidated through mgt tools	Sys monitors correlates & recommends actions	Sys monitors correlates & takes action	Components dynamically respond to bus policies
Skills	*Extensive, highly skilled* IT staff	IT staff *analyzes & takes actions*	IT staff *approves & initiates actions*	IT staff *manages performance* against SLAs	IT staff *focuses* on enabling business needs
Benefits	Basic Requirements Met	Greater system awareness Improved productivity	Less need for deep skills Faster/better decision making	Human/system interaction IT agility & resiliency	Business policy drives IT mgt Business agility and resiliency

Manual → **Autonomic**

aspector@us.ibm.com

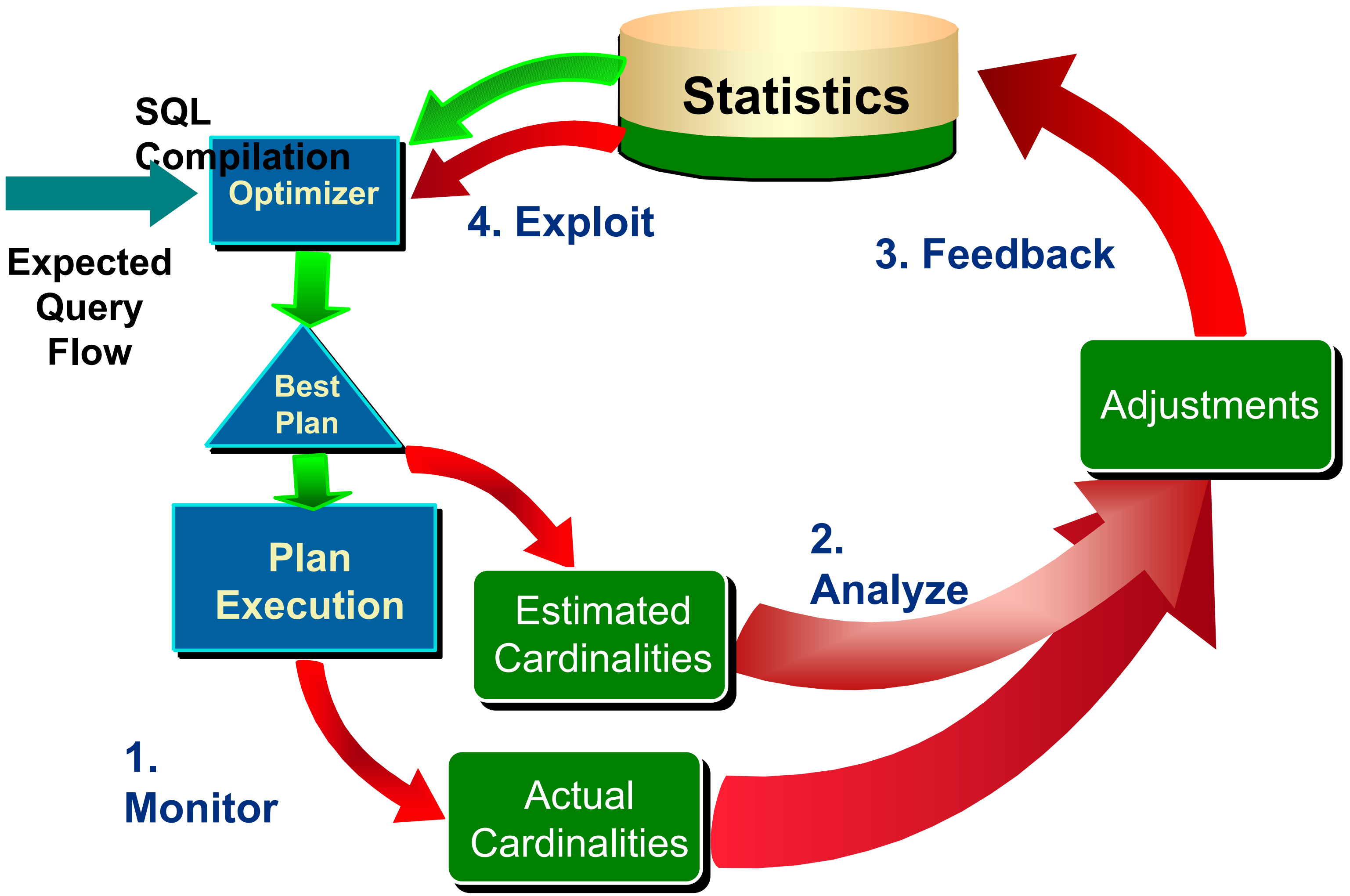
IBM®

Autonomic Computing Res. Projects



aspector@us.ibm.com

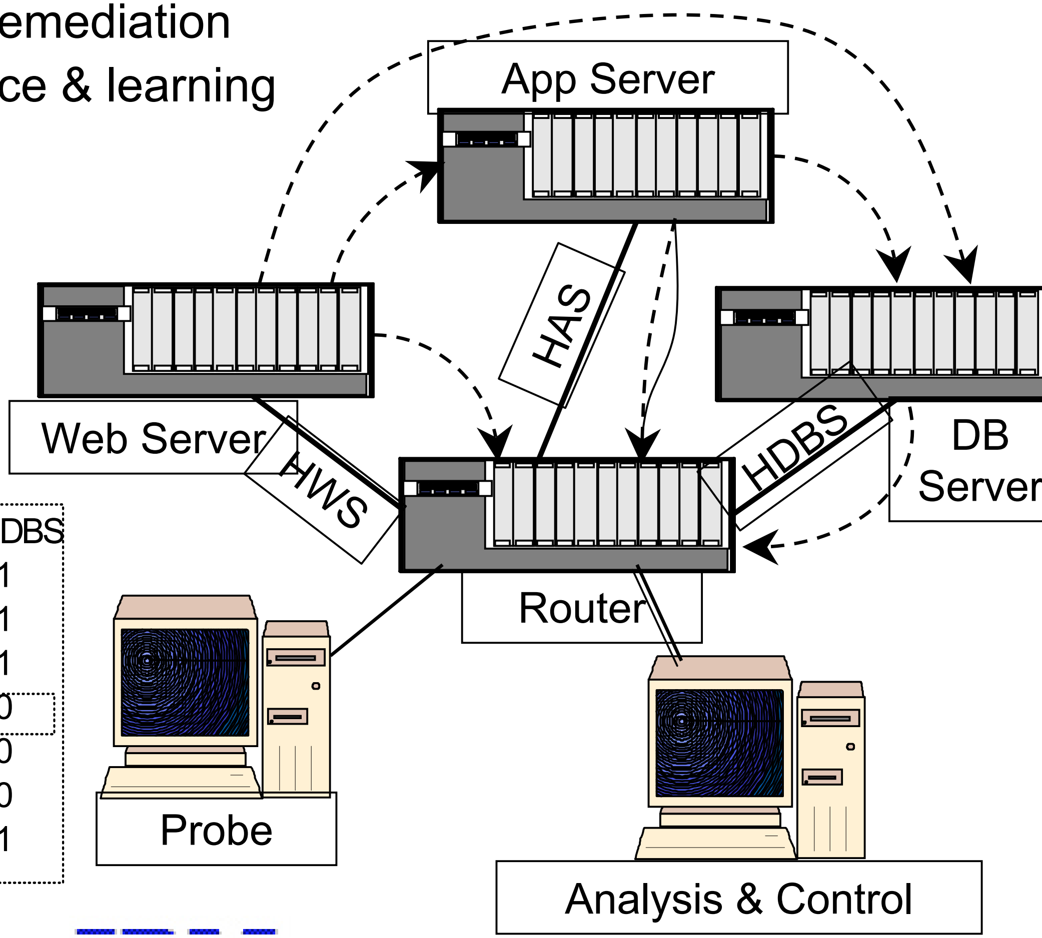
LEarning Optimizer for DB2 (LEO)



aspector@us.ibm.com

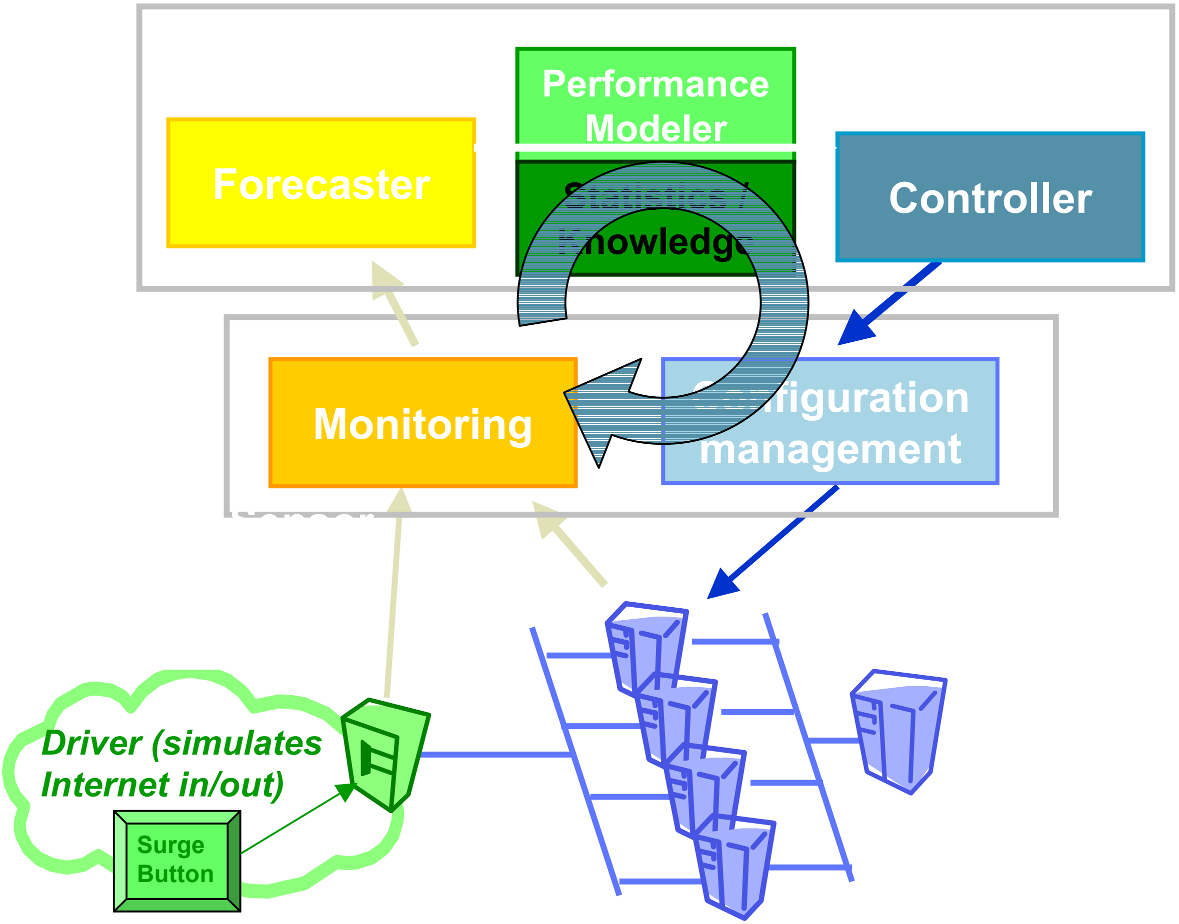
Dependency Mgt & Problem Determination

- Determine functional dependencies among elements
  - Mine design docs, system config metadata, log files
  - End-to-end probe platform for running system
- Use dependency information for system management
  - Problem localization & remediation
  - Real-time active inference & learning



	WS	AS	DBS	R	HWS	HAS	HDBS
pWS	1	1	1	1	1	1	1
pAS	0	1	1	1	0	1	1
pDBS	0	0	1	1	0	0	1
pingR	0	0	0	1	0	0	0
pingWS	0	0	0	1	1	0	0
pingAS	0	0	0	1	0	1	0
pingDBS	0	0	0	1	0	0	1

Adaptive Workload Surge Protection

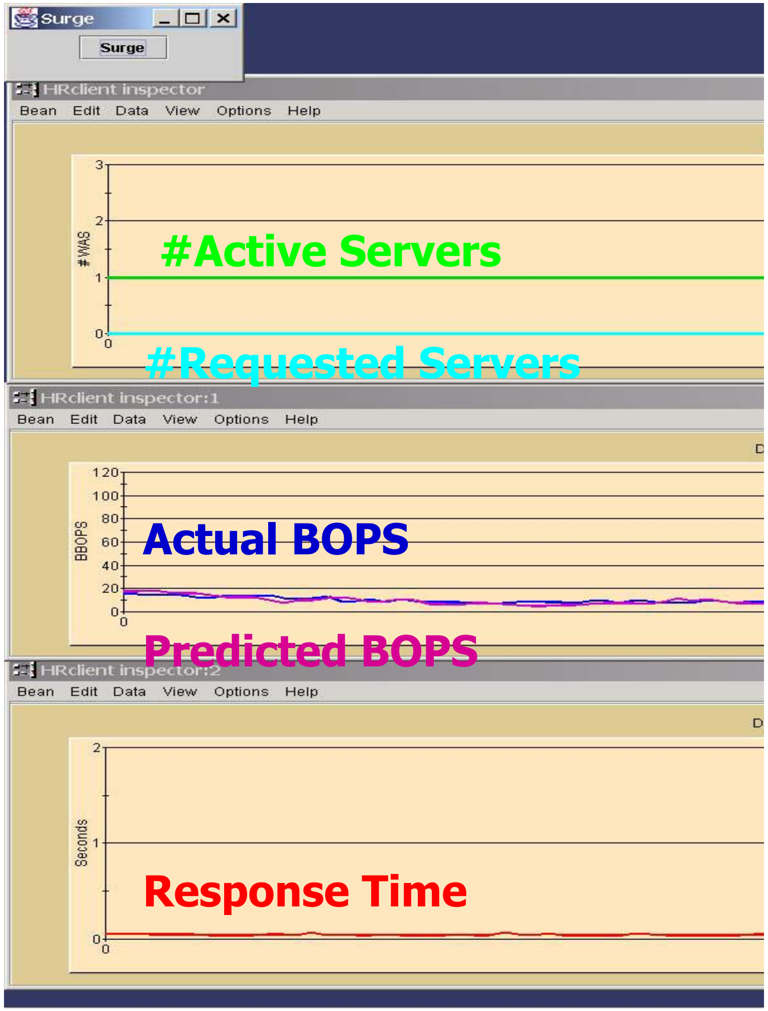


aspector@us.ibm.com

6/17/03

IBM

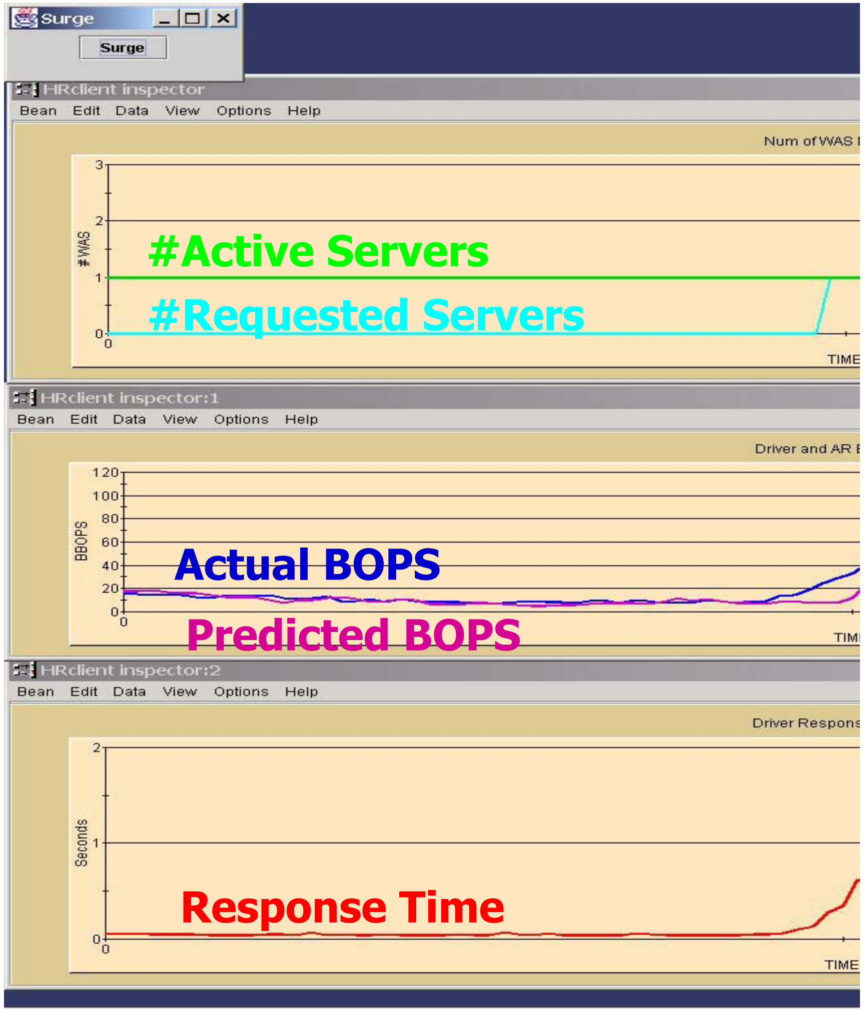
Surge Protection Demo – *Steady State*



aspector@us.ibm.com

IBM®

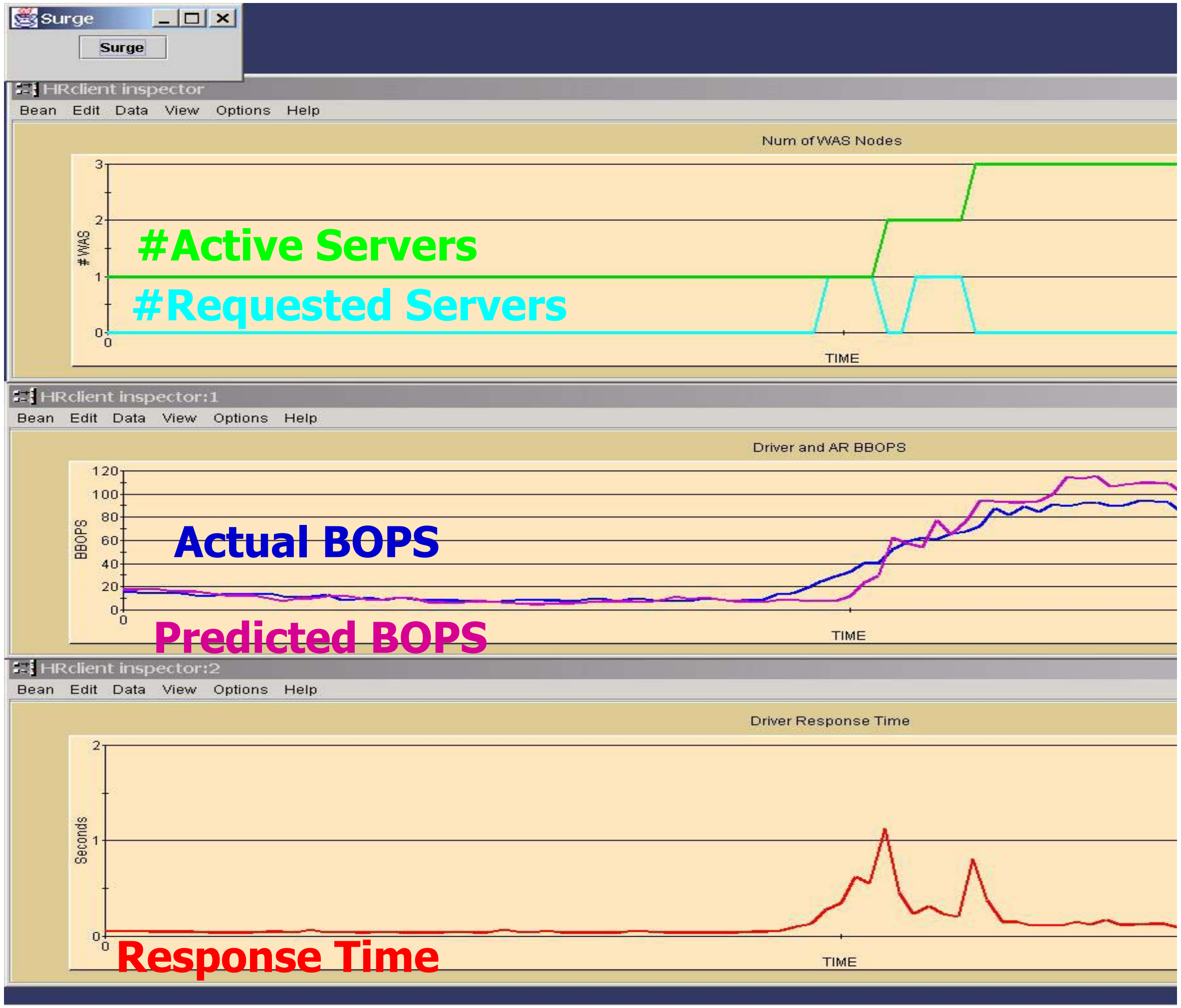
Surge Protection Demo – *Surge onset*



aspector@us.ibm.com

6/17/03

Surge Protection Demo
Forecast Surge and Provision Servers

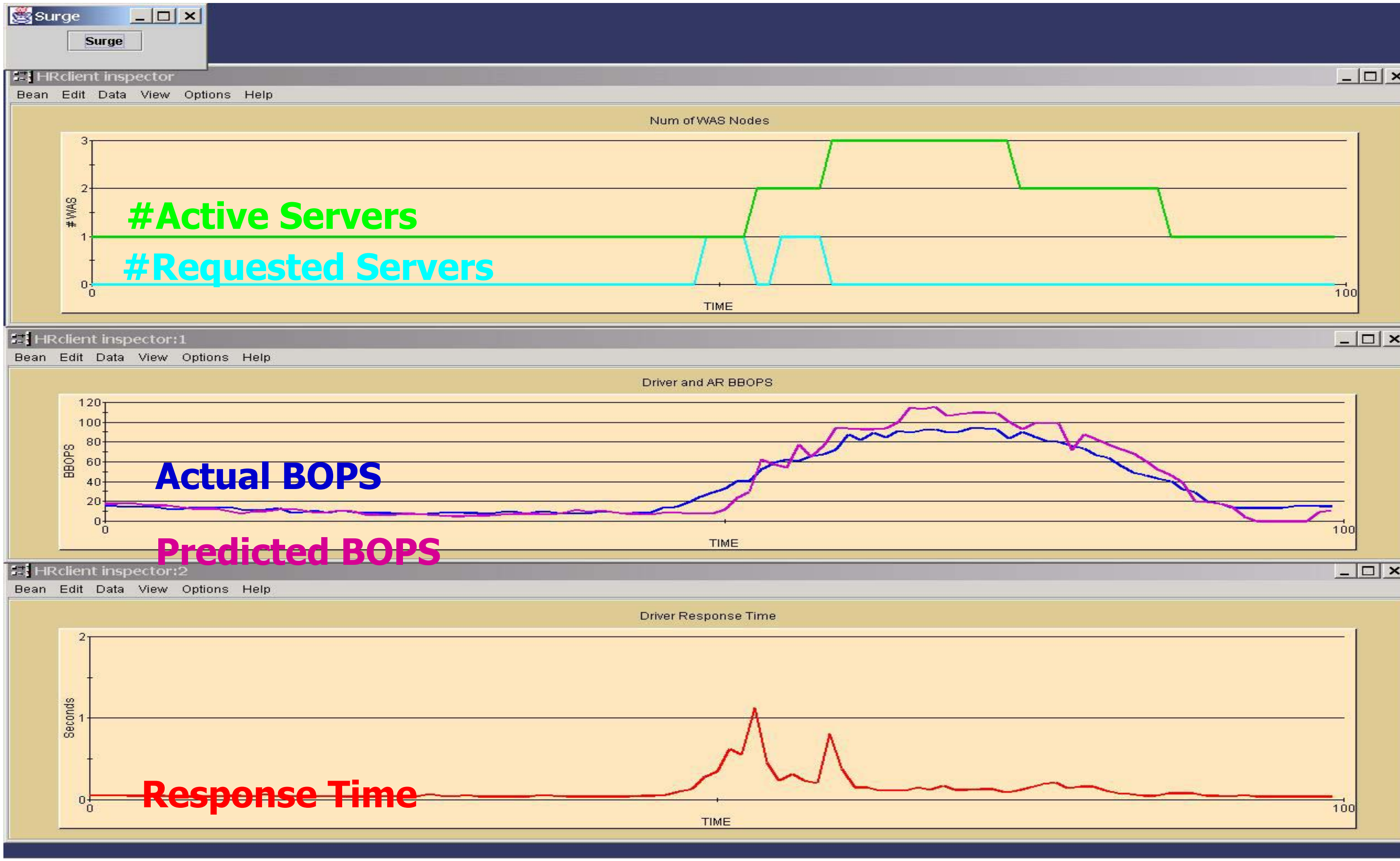


aspector@us.ibm.com

6/17/03

IBM®

Surge Protection Demo
Monitor and Remove Servers

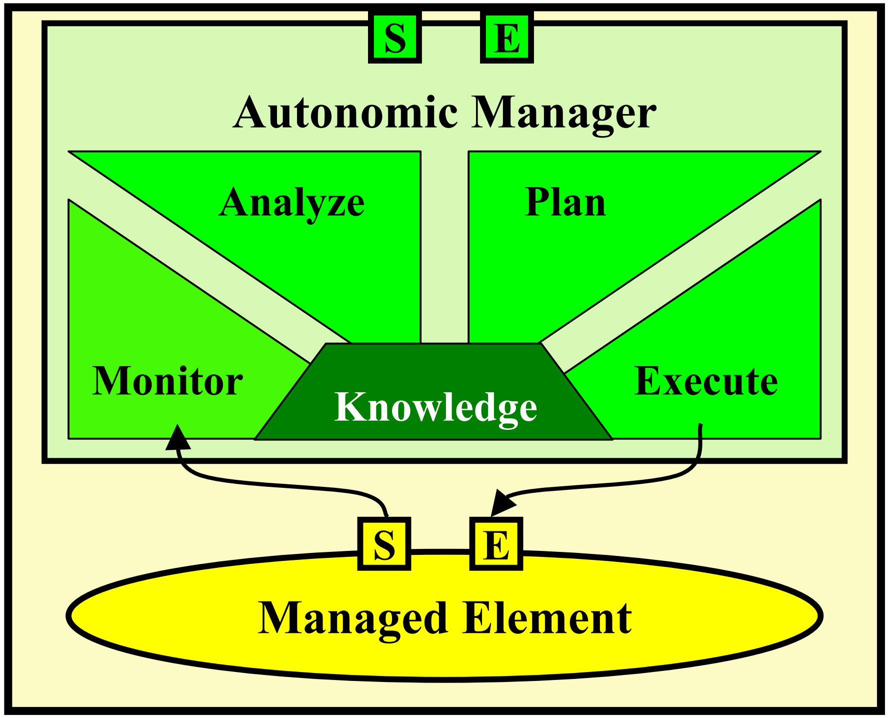


aspector@us.ibm.com

IBM

Autonomic Computing Architecture – 1.0

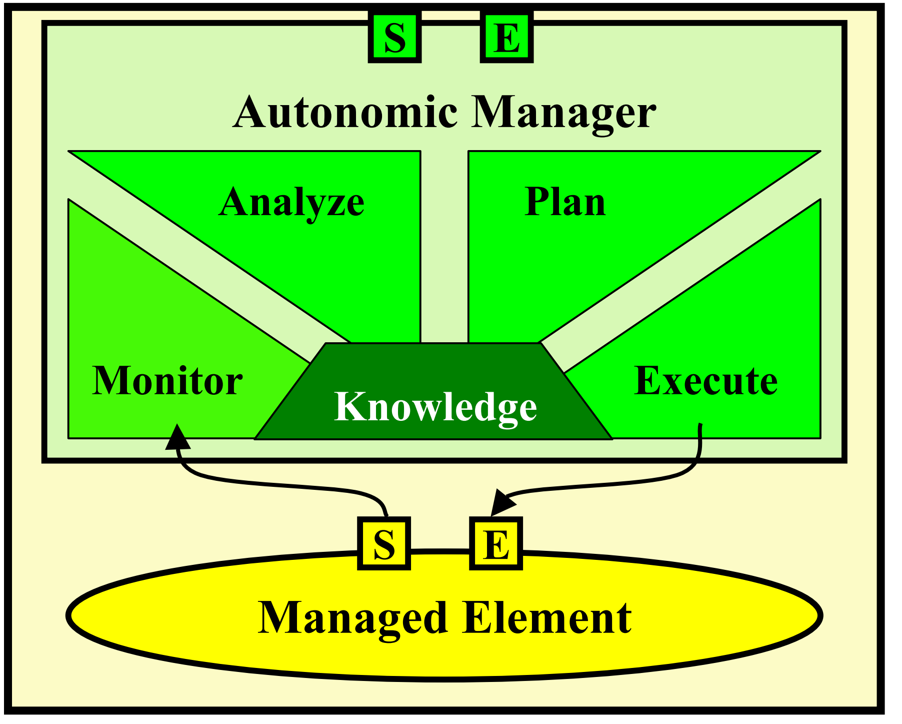
- Managed element – provides function
  - Basic resource - database, storage system, server, software app
  - Higher level - manages other elements
- Autonomic manager – provides mgt
  - Monitor – gather data
  - Analyze – assess performance
  - Plan – determine response
  - Execute – implement response
  - Knowledge – state & policies
- Autonomic element
  - Provides/consumes services
  - Interacts w/ other autonomic elements
  - Manages in accordance w/ policies



An Autonomic Element

Autonomic Manager Toolkit

- Facilitates autonomic mgr construction
  - In accordance w/ AC architecture
- Catcher for generic AM technologies
  - OGSA messaging
  - Policy tools
  - Monitoring technologies
  - AI tools for knowledge representation, reasoning
  - Math libraries for modeling, analysis, planning
  - Feedback control
- V1.0 soon available publicly

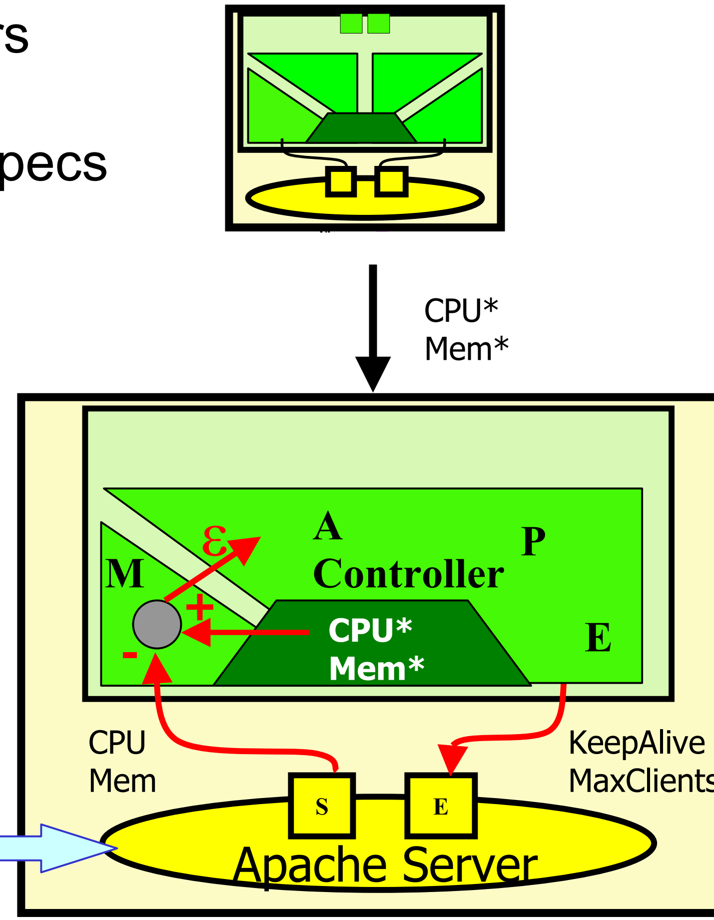


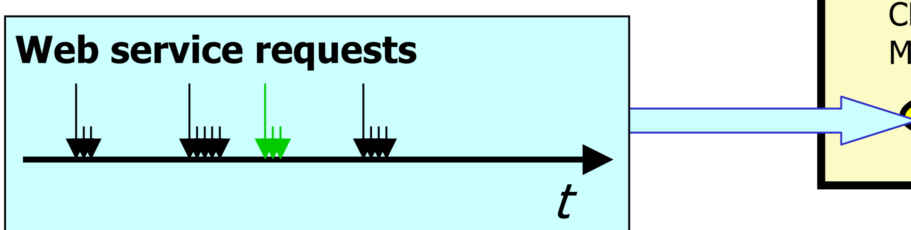
An Autonomic Element

aspector@us.ibm.com

Generic Adaptive Control

- Feedback control to tune effectors
- Based on high-level behavioral specs
  - Multiple goals
  - Multiple effectors
  - Time-varying demand
- Various database and server applications





aspector@us.ibm.com

Summary – Autonomic Computing Research

- Ad hoc autonomic elements
  - Many valuable prototypes & products built
  - Important lessons learned
  - Continuing research interest
- Ad hoc autonomic systems
  - Some research prototypes functioning
  - Early results valuable
  - Continuing to evolve and improve
- Structured autonomic elements
  - Version 1.0 complete
  - Substantive reviews underway
- Structured autonomic elements
  - Early projects underway
  - Architecture in formative stages

ISR Annual Research Review

Continual Optimization

aspector@us.ibm.com

6/17/03

IBM®

Continual Optimization

- Almost every resource can almost always be connected
- Most resources' usage can therefore be monitored or changed
- The opportunity for optimization is great
- Continual optimization could fundamentally change how we might lead our lives

ISR Annual Research Review

Continual Optimization

Just as Computing has Evolved

batch processing	individual transactions
ab initio processes	state based processes
scheduled computations	as needed, triggered by events
relatively expensive	virtually free
independent processes	interacting processes
back office	end users
rigid linear workflow	adaptive, state-based workflow
islands of computing	global networking

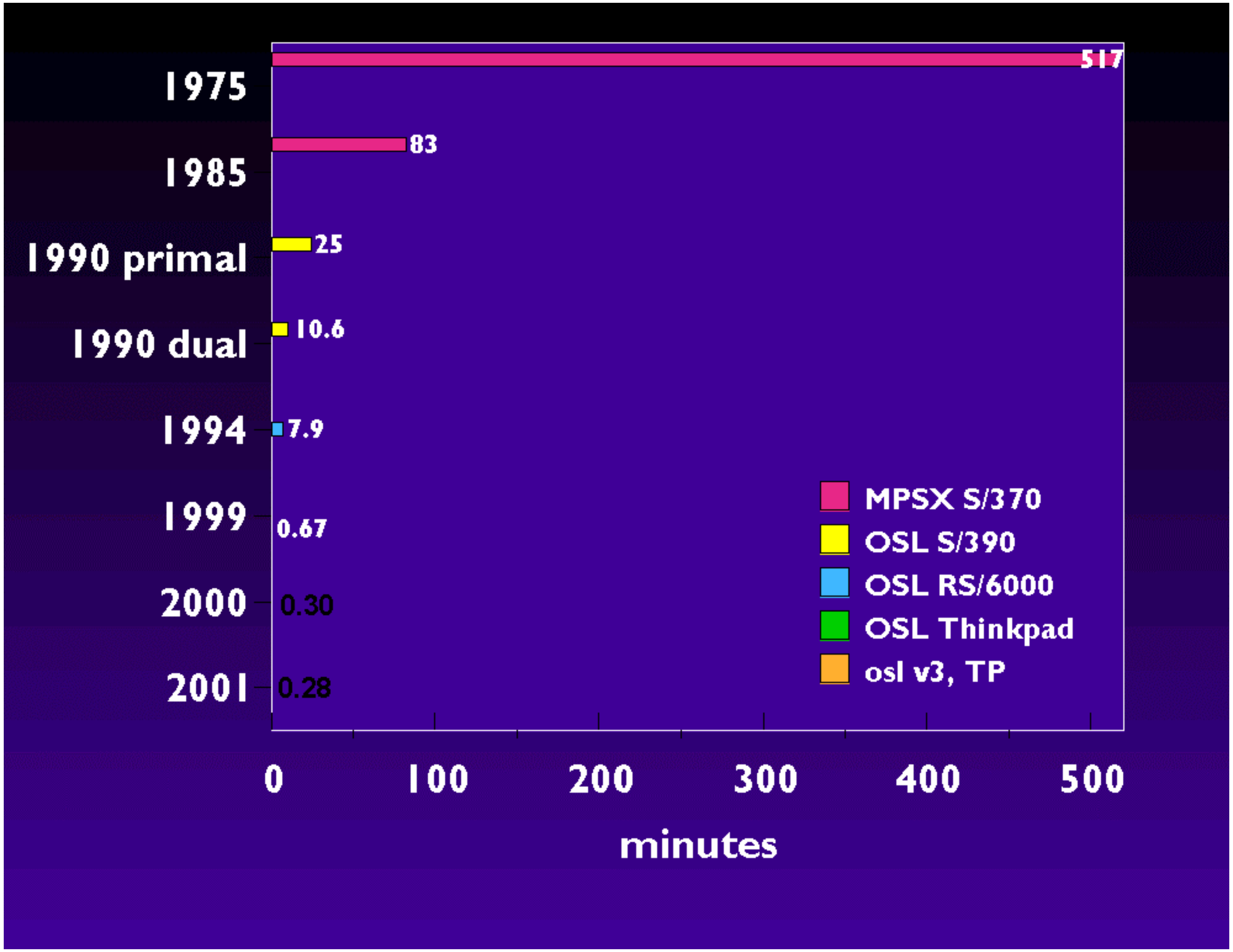
Continual Optimization

Analytic Computing is Also Evolving

enabling new business applications and business models

large scale analysis for strategic decisions	mixed scale analysis for planning and operational decisions	continuous planning and replenishment in supply chain
ab initio models and algorithm	state based analysis, to update and adjust plans	personalized B2C web pages
scheduled analysis with published plan	analysis as needed, triggered by events, only next actions communicated	mobile technician scheduling and dispatching
relatively expensive	virtually free	complex marketplaces and auctions
independent static models on separate data	interacting dynamic models, sharing common data and interfaces	replane: airline seat reallocation
analytic experts in back offices	real-time analysis by end users or agents	fuel optimization for truckers
deterministic,models with single proven algorithm	stochastic models that are solved by modifiable algorithms	e-commerce fraud detection
data from isolated, static database	data gathered online from arrays of networked sensors	smart freight instrumented power grid

United Airlines 8 Fleet Problem



Airline Optimization

Planning Horizon

Hour

Short Range Yield Mgt
e.g. auctions
(Priceline)

DYNAMIC/ADAPTIVE
Resource and
Demand management

Current Weather

Day

MANUAL
- Flight schedule repair
- Flight re-routing
- Crew schedule repair

LOCAL OPTIMIZATION
- Flight schedule repair
- Flight re-routing
- Crew schedule repair

GLOBAL OPTIMIZATION
- Flight schedule repair
- Flight re-routing
- Crew schedule repair

Month

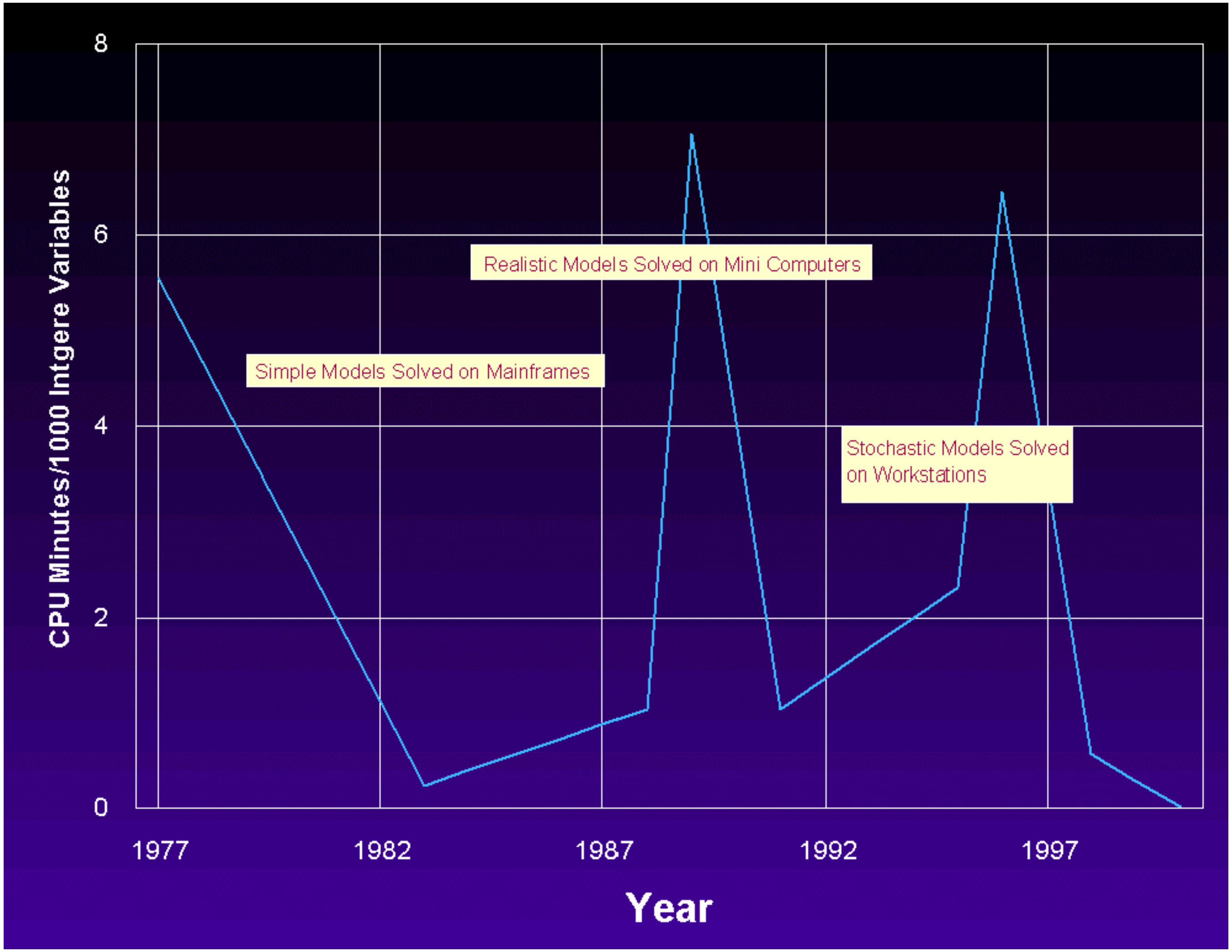
LOCAL OPTIMIZATION
- Flight schedules
- Fleet assignments
- Plane assignments
- Crew schedules

GLOBAL OPTIMIZATION
- Flight schedules
- Fleet assignments
- Plane assignments

GLOBAL OPTIMIZATION
- Flight schedules
- Fleet assignments
- Plane assignments
- Crew schedules

1990 **2000** **2010**

Unit Commitment Problem



Optimization Examples

- Season tickets in advance
- Pricing set above MC and to clear the market
- Static binding of resources
- Approximate production optimization
- Opportunistic interpersonal scheduling
- Search for a restaurant while driving on the road

- Notification when you want of events you like
- Pricing based on utility of consumer
- Dynamic resource binding
- Exact Production optimization
- Dynamic interpersonal scheduling
- Be informed of nearby restaurants meeting criteria

Dinner and a Show in the Old Economy

ISR Annual Research Review

- Consumers plan ahead, but also react
  - Order Show tickets by phone for box office pickup
  - Phone for restaurant reservations
  - Plan travel route using static map data
  - React to traffic, parking shortages
- Providers: Allocate and sell capacity
  - Pre-allocate blocks of tickets to channels
  - Sell specific seats
  - Allocate tables in advance or FCFS with little care
  - Sell FCFS
  - Broadcast traffic information

aspector@us.ibm.com

IBM®

Dinner and a Show in the e-Economy

- Consumers: Buy online
  - Order Show tickets by internet, print at home
  - Internet reservations, updates via wireless
  - Determine travel route using current traffic conditions
  - Re-optimize manually in response to traffic, parking shortages

- Providers: Allocate & sell capacity - but closer to the time of use
  - Manage inventory across multiple channel
  - Sell specific seats
  - Allocate tables
  - Maintain dynamic traffic data
  - Sell FCFS

Dinner and a Show w/Continual Optimization

- Consumers: Express preferences, receive complete offers
  - Show, dinner reservations, travel route and parking for a single price
  - All can be reserved ahead of time
  - Special last minute deals available
  - Monitored and re-optimized dynamically (according to customer preferences, of course)
- Providers: Dynamically allocate and aggregate
  - Allocate multiple resources consistently
  - Dynamically manage inventory across multiple channel
  - Sell seats/tables in categories & do late binding of specific seats
  - Supply reservations for parking, travel lanes, etc.
  - Monitor availability & re-deploy dynamically in response to disruptions
  - Slack management

A Real Example: Limo Scheduling & Dispatching

- Executive class ground transportation
- Service by owned resources in 7 cities and worldwide through affiliates
- Pride in excellent service record of 98-99% on-time pickups; but at cost of 10-12% request refusal rate at peak times and low utilization of resources
- Customer contacted IBM Research through IBM Innovation Center
- Scheduling problem recognized as a potential match for Continual Optimization initiative
- Integration, middleware, project management, etc. from IBM Business Consulting Services

Project Overview

- Watson Optimization Center developed a Continual Optimization scheduling/dispatching tool for LimoCo
  - Optimizer code delivered in 2Q02
  - Live tests of the system in December
- Project size
  - 5600 lines of custom code
  - + existing optimization libraries
  - + databases, integration and user interfaces by BCS
- Success criteria
  - Current driver utilization ~10% below optimal, believed to cost 30-100M/year.
  - Our solution within 1.5% of optimal (offline); online unknown

Some Details

- Three modes for optimization of driver/vehicle schedules
  - daily "offline"
  - 15 minute "continual"
  - 15 second "instant" on-demand mode
- Minimize costs
- Constraints
  - Schedule all rides if possible (remainder subcontracted)
  - Be on time whenever possible
  - Send licensed vehicle/driver on airport trips, if applicable
  - Send best drivers to VIP passengers
  - Anything else customer dreams up
- Monetary (funny money) penalty if constraint not met
- Multiple depots; some with limited number of vehicles
- All outputs advisory only; dispatcher has final say
- Per city: ~1000 reservations/day, 200 drivers/150cars

Solution Overview

- Offline and continual modes modelled as integer program
  - IP solved using Optimization Solutions and Library (OSL)
  - Within 1.5% of optimal on 600-ride instance in <5 minutes
  - In continual mode, some variables are fixed; time limits are imposed
- Iterative improvement in instant mode
  - deterministic and randomized hill-climbing heuristics
  - basic idea: try to put unscheduled ride on each driver's schedule; choose cheapest alternative
    - more sophisticated versions
  - repair schedule after user overrides, reservation & real-time updates
- Iterative improvement also used for consistency maintenance
  - take advantage of solution to next continual run
  - with user overrides
    - user's say is final
  - in response to data updates
  - may make current schedule infeasible
- This is a non-trivial part of Continual Optimizationb

aspector@us.ibm.com

ISR Annual Research Review

Issues: Communications & Data Infrastructure

- Data
- Transaction management
- Modular, distributed architecture
- Privacy
- Availability
- Scale
- Organizational autonomy
- Ease of use pervasive devices | HCI
- Most significant problem: *Business Process Integration*

Enormous Question

- How successful will we be with business process modeling and automatic generation of the code
- From strategy to implementation?

Dynamic Allocation Technology

- General questions
  - How dynamic?
  - How optimal?
  - How valuable?
- Allocate resources to customers
- Allocate revenue to providers
- Update allocation in response to changing conditions

aspector@us.ibm.com

IBM®

New Applications of Known Models/Algorithms

Shortest path problem (single car routing, static network)

- Easy, scales almost linearly with network size O(n+m) log m

Time dependent shortest path (speed depends on time of day)

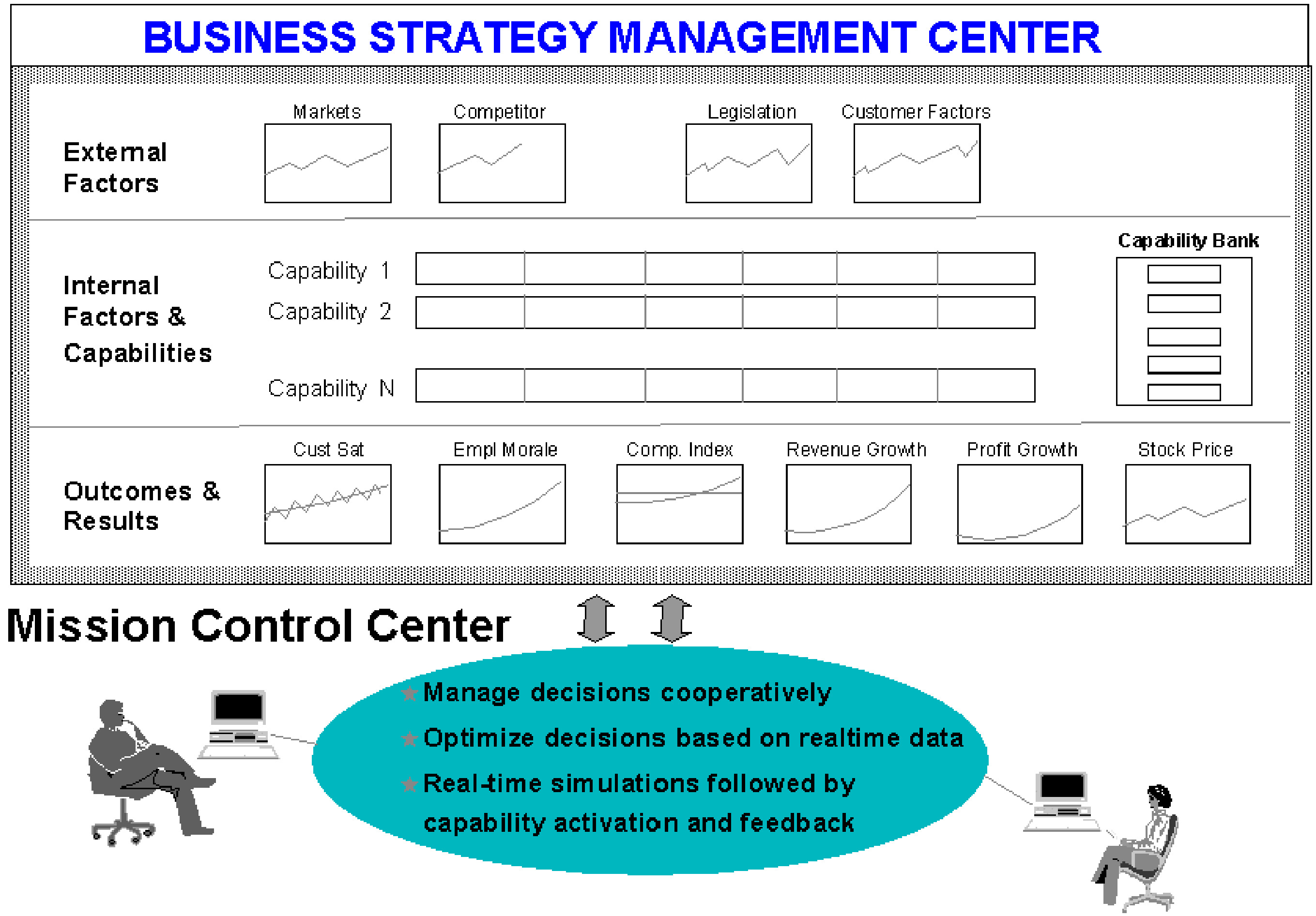
- Expand network to represent time intervals
- Deterministic or Probabilistic
- Dynamic updates based on state

Multiple vehicles with multiple origin-destinations

- capacity limits on roads: multi-commodity network flow model: NP-comp
  - Routinely solved for 1000's of links and cars
- Incremental approaches (shortest path subj. to residual capacity) work well
- Integer programming using a column (path) generation approach appears to scale well (1000's of roads and vehicles)
  - adjust speeds according to congestion and resolve

Other resources (parking, restaurant) can use job shop scheduling. Many jobs & machines, but nice structure & few ops/ job

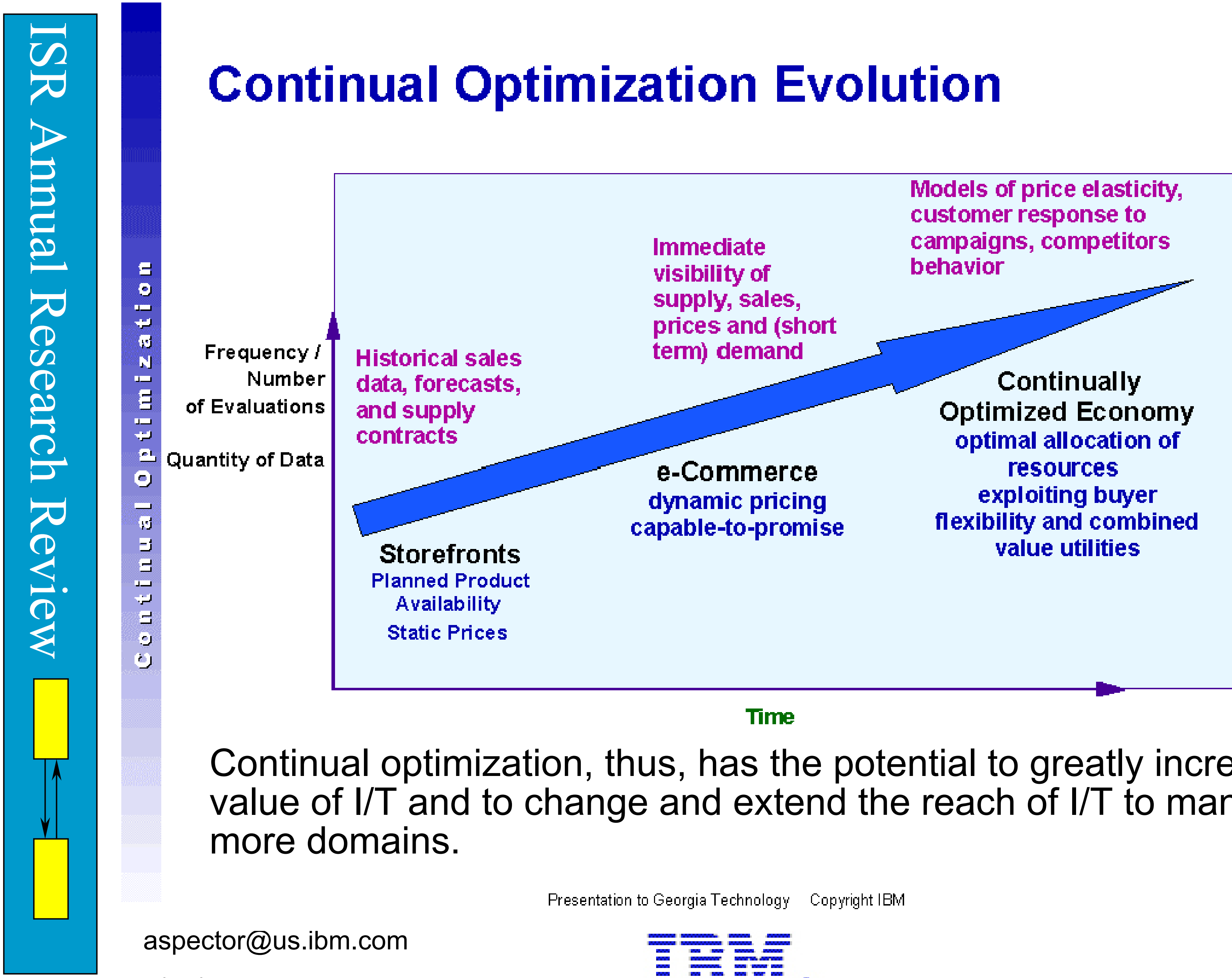
Use of simulation & optimization tools and "active real-time" data for strategic business decisions.



aspector@us.ibm.com

IBM®

Continual Optimization Evolution



Continual optimization, thus, has the potential to greatly increase value of I/T and to change and extend the reach of I/T to many more domains.

aspector@us.ibm.com

6/17/03

Summary and Conclusions

aspector@us.ibm.com

Autonomic Computing

- Subsystem design improved to eliminate manual control
- Core techniques:
  – Control theory
  – Increased use of rules systems; perhaps, with inference & common sense
  – Negotiation
- Standardization of event reporting to provide opportunities for data mining, statistical machine learning, and more feedback control
- Architecture

Continual Optimization

- The connectivity is there
- The transaction costs are there
- The mathematical methods are substantially there
- Can we get over the complexity issues of Business Process Integration to achieve enormous benefit to our field and potentially to society?

Conclusion

- Computing can provide greatly increased value to society
  - But, we must conquer complexity
  - And do more than automate what we previously did manually
- This talk illustrated two important sub-problems
- But, there is much more to be done to unlock the value of these wonderful machines without undue complexity and cost