

WAN provisioning for application deployment: A practical approach

A major role for IT staffs is timely support of new business initiatives. IT staffs don't want to be the bottleneck in the application deployment process, yet they can't afford to put a new application into the network and have it fail to perform. Few have unlimited resources to staff a specialized capacity planning group for application deployment; and even fewer (if any) can afford to throw away corporate dollars and over-build their network "just in case."

This Catch-22 describes a very real problem most IT staffs face. The modeling-based capacity planning solutions pitched by vendors have missed the mark, both in the technology and in its implementation. While occasional successes are possible by applying brute force, achieving a consistent, repeatable process that can demonstrate an ongoing ROI is not. In the words of one disgruntled but representative user, "Each time I used the solution, I felt like the vendor owed *me* money."

In this white paper, we'll review the specific goals and functions of rapid application deployment (RAD) and how network capacity planning fits in the process. We'll discuss the ROI based on just-in-time network design and how capacity planning promises to deliver this. Then we'll look at the technological and organizational hurdles that have caused so many vendors' solutions to fail. Finally, we'll take an enlightened look at a well-defined solution that supports the goals of RAD and avoids the all-too-well-known pitfalls.

Defining rapid application deployment

The goals of rapid application deployment are universal and familiar to most IT organizations today. To overcome the time-to-market pressures of new business initiatives, an enterprise must be able to deploy a new application into a production environment as quickly as possible, reliably and with as little risk as possible.

Of course, these are somewhat conflicting goals. Speed often increases risk, and no significant deployment can be entirely without risk (nor should it be, since that would mean you are spending too much time testing).

Deploying a new application into a production network encompasses multiple disciplines, often involving teams that have not worked together (or worked well) in the past. These teams will certainly have different perspectives on technology; application development, server capacity analysis and network design require quite different skill sets. These teams may even have conflicting business goals. On the business side, the application developers may be working for a business unit whose goal is to increase revenue. On the operations side, the network planners will be working for a business unit whose goal may be to control costs.

This is not to say that these conflicting goals are mutually exclusive—but to accommodate both requires a level of communication and cooperation that has only recently seen a significant degree of success. These successes have been predicated on a maturing rapid application deployment process enabled by cross-functional cooperation within the organization and tools that deliver meaningful measurement and analysis of performance from multiple vantage points.

Rapid application deployment is ideally delivered as a loosely integrated series of “best practices” that test critical components of the system in the context of the intended deployment environment. Some of this involves measurement, some of it projection. Core components of the solution are application profiling, load testing and network capacity planning.

- **Application profiling** analyzes the impact that the intended production network environment will have on the performance of the application, delivering quality-of-service requirements to the network planner as well as “networkability” tuning input to development teams.
- **Load testing** analyzes server performance under various load conditions, providing server configuration and sizing information as well as pre-deployment bottleneck analysis.
- **Network capacity planning** analyzes the impact that the additional application traffic will have on the performance of the network, and should focus on both WAN link sizing and end-user response time.

(The benefits and implementation of rapid application deployment are described in more detail in the white paper “Taking the Risk Out of Rapid Application Deployment,” available from your Compuware representative.)

Since the focus of this paper is on network capacity planning as part of a rapid application deployment solution, we should first understand the cost justification in order to focus our efforts.

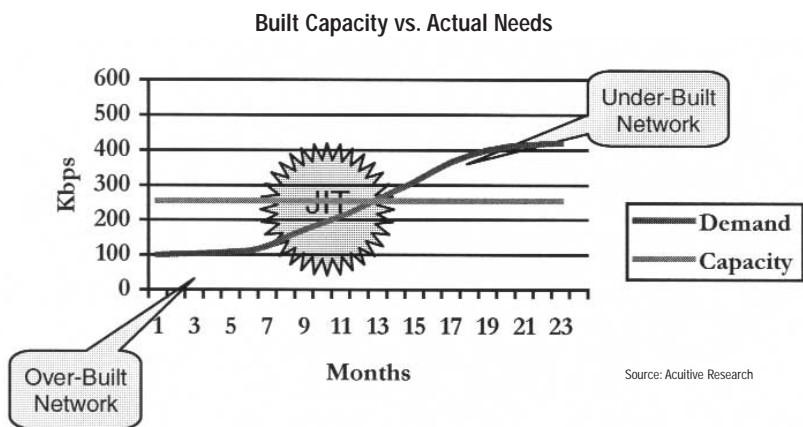
Just-in-time network design

The principles of just-in-time (JIT) network design, according to Acuitive, Inc., a research and start-up acceleration consultancy, are simple, practical and achievable. Most of the time IT organizations spend trying to anticipate the future is wasted. With a flexible network design, good communication between groups and reasonable instrumentation, there are two key areas where expending planning effort is consistently worthwhile—trending and new application deployment. Each of these can be clearly and specifically identified, quantified and projected. If the infrastructure focus is on WAN bandwidth, typically the most significant part of the network budget, the planning ROI can be easily identified (the white paper “Just-In-Time Network Design” is available from Acuitive, Inc. at <http://www.acuitive.com>).

To some degree, WAN links often operate in one of two undesirable states:

They are over-provisioned. Anticipating an as-yet undefined demand for greater bandwidth, you have allocated budget to provide more “headroom” than you really need. While you are technically prepared, you are also wasting money. Without a true rapid application deployment process in place, of course, you probably don’t have much warning of change driven by new application deployments. You have chosen to mitigate some of the risk of application deployment failure, and also to avoid becoming the bottleneck for speedy delivery. But there is most definitely a cost associated with this approach—not necessarily to the business unit, but a recurring infrastructure cost.

They are under-provisioned. Unable to anticipate specific bandwidth requirements, you have selected cost-control as your guiding principle, running a tight network and squeezing out as much return on bandwidth investment as possible. Without a true rapid application deployment solution in place, you will wait for the application pilot or early deployment phase to see what happens. The likelihood is, of course, that the next new application to be deployed into the network will perform poorly, also causing existing applications to experience slowdowns. There is also a cost associated with this approach; not to the infrastructure this time, but to the business units the infrastructure supports, measured in lost opportunity and reduced productivity.



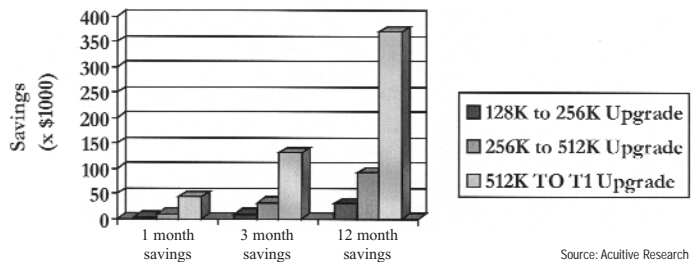
Source: Acuitive Research

Figure 1: The just-in-time zone

Just-in-time's ROI

Gartner estimates that a corporation with 73 sites will spend roughly \$658,000 per year in carrier WAN costs alone. While “perfect” JIT network planning would provision bandwidth coincident with demand, factors such as variable carrier installation lead times, incremental bandwidth tariffs and “play” in the project often dictate a degree of discretion. Using the example of the corporation with 73 WAN links, the cost of over-provisioning becomes clear.

WAN Savings with JIT Deployment (Corporations with 73 WAN links)

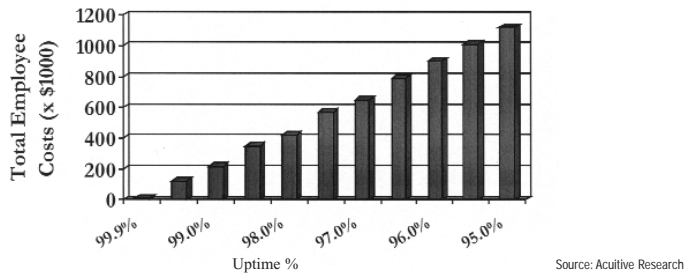


Source: Acuitive Research

Figure 2: Just-in-time WAN savings

Similarly, an under-provisioned network extracts its own cost to the corporation. Lost productivity and squandered business opportunities resulting from poorly performing applications and overloaded networks have an equally significant impact on the corporation's bottom line.

Productivity Costs of Network Downtime (5000 user network)



Source: Acuitive Research

Figure 3: Lost productivity costs

These are illustrations of the cost benefits related directly to accurately allocating WAN bandwidth; ROI based on application profiling and load testing is separate and cumulative, and the integration of these components into a rapid application deployment solution compounds their individual benefits.

WAN capacity planning: Key to rapid application deployment

If the value of a rapid application deployment solution is clear, then the importance of effective WAN planning should also be evident. But isn't this just a subset or function of network capacity planning? Isn't network capacity planning an entrenched business practice? Doesn't this already deliver planning information for WAN links in preparation for new application deployments? The answers are yes, maybe and not effectively.

Capacity planning is essentially a process of anticipating the future; from a network perspective, this means predicting the impact that change will have on the quality of service (QOS) the network will deliver, often reduced to bandwidth requirements. Business change may be driven by a new initiative that will be automated or serviced by a new application (the focus of this document); it may also be driven by growth (positive or negative) of an existing business, by physical expansion or consolidation of corporate offices, or by mergers and divestitures. Technology change may also demand capacity planning, such as the implementation of new carrier services, server consolidations or network reconfigurations.

Network capacity planning is therefore an extremely complex discipline, and the infrastructure to support this function quite complicated. The modeling tools used for capacity planning and for analyzing the impact of these changes attempt to address all of the cases listed above (and many more), and as a result have become particularly complex and unwieldy. Effective use requires extensive (and expensive) modeling expertise, including in-depth knowledge of network technologies, protocols and distributed application architectures, as well as modeling theory, techniques and approaches.

Building software models to support these diverse goals requires gathering detailed topology information, either defined manually or perhaps discovered, imported and cleaned up. The traffic portion of the model requires extensive and interpreted data from network monitors. More often than not, the resulting model bears little resemblance to reality.

While there are some reasonable applications for these modeling tools, they are in fact an inappropriate approach to most important capacity and performance questions. Applying these as part of a mature RAD solution does not work, since the resulting delay to the process is unacceptable in view of the desire for speed. The result is that most applications are deployed with virtually no network capacity planning.

WAN provisioning defined

We have seen that anticipating the deployment of a new application is one of the two primary business changes for which applying a capacity planning process might be justified. WAN provisioning can be defined as the specialized process of identifying WAN bandwidth requirements to support new and/or revised application deployments. WAN provisioning is important because it:

- focuses on the most critical network performance bottleneck—the LAN/WAN interface, where speed differences often are an order of magnitude or greater
- relies on readily available, verifiable input data
- delivers actionable information that can be directly related to infrastructure costs.

In order for WAN provisioning to support the goals of rapid application deployment, it must first overcome the organizational and technological hurdles of traditional capacity planning. In other words, it must not require extensive modeling expertise, it must not require complex modeling tools, and it must not rely on hard-to-obtain input data. To this end, WAN provisioning leverages the application profiling process and provides a tightly coupled solution focused on WAN bandwidth requirements. The result is that WAN provisioning is easy to implement and scalable according to the scope of the deployment requirements.

WAN provisioning details

Application profiling answers the questions “What impact will the network’s QOS have on my application’s performance?” and “What can I do (to the application, server or network) to improve this performance?” WAN provisioning answers the complementary questions of “What impact will the application’s traffic have on existing network utilization?” and “How much bandwidth will be needed to support a specific response time service level?”

To answer these questions, WAN provisioning relies on four key profiles:

The application profile—the traffic and performance characteristics of the key end-user transactions that make up the application. The application profile describes the network conversations that occur between each tier as well as the processing delays that occur at each node. The application profile is created during the application profiling part of the rapid application deployment process, so it is a readily available input to WAN provisioning. It is important to note that the application profile is real, measured data, so the margin of error inherent in this profile is minimal.

The user profile—the description of how users will interact with the application. The user profile describes the frequency and mix of those key transactions measured in the application profile. The user profile should first be defined at the business process level, where the number of business functions (such as new reservations made, orders taken, accounts entered) per hour is defined. Once the business functions have been identified, they are mapped to the underlying application transactions in the application profile. Since user profile definition usually relies on interviews of the department manager and application users, it likely represents the largest margin of error, so a little extra effort applied to this part of the process will be important. Sometimes, two sets of profiles are created; one representing the anticipated use, another representing a “worst case” scenario.

The deployment profile—the intended locations and numbers of clients and servers, as well as the time of day being analyzed (usually the “busy period”). The deployment profile maps user profiles to physical sites, and corresponds to the phased implementation plan for the application itself.

The infrastructure profile—the definition of how remote user locations connect to the server location(s). The infrastructure profile describes physical and logical WAN links, and often will also include existing link bandwidth, background loads and latencies. Since the focus of WAN provisioning is on WAN bandwidth, complex LAN topologies and interconnect devices such as switches are not required.

WAN provisioning will help define how much WAN bandwidth will be needed to support the application deployment described by these profiles, by providing accurate and meaningful business decision metrics. WAN provisioning delivers this in a multi-dimensional, interactive report that illustrates the relationships between bandwidth, utilization and end-user response time. Tradeoffs between response time (i.e., productivity) and bandwidth (i.e., infrastructure costs) can be quickly evaluated as they relate to the business deployment plan, enabling true just-in-time network design.

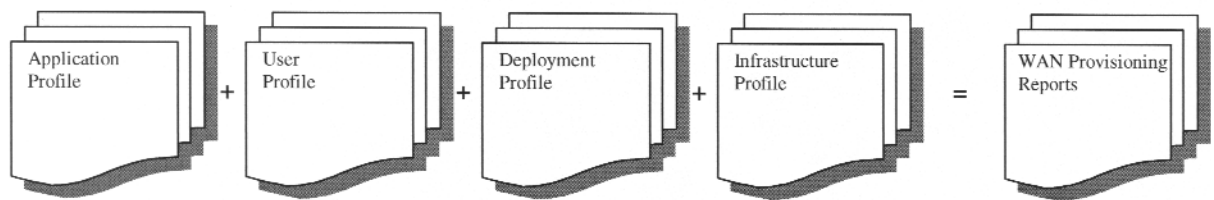


Figure 4: The WAN provisioning process.

Levels of WAN provisioning

All application deployments are not created equal; some may be quite small, supporting individual users at only a few sites. Others may be quite large, supporting thousands of users worldwide. While most fall somewhere in between, the WAN provisioning process should be flexible enough to support the range of possibilities without imposing the use of a single tool, yet consistent enough that it can be applied to virtually all cases. We can define three “levels” of WAN provisioning to support this. In all cases, Compuware’s Application Expert is used to create the application profile used as input to WAN provisioning.

The transaction level—For single users of an application at individual sites, or for multiple occasional users where concurrent use is not anticipated, WAN provisioning is actually part of the application profiling process. Using Application Expert’s sophisticated Response Time Predictor (RTP) feature, recommendations for bandwidth can be evaluated simply in terms of end-user response time. There is no real user profile required, because a single user will only run transactions sequentially. The deployment and infrastructure profiles become the inputs to the RTP, defining bandwidth, latency and background load for the network connections between each tier or node in the transaction. The Application Expert user is already performing transaction-level WAN provisioning.

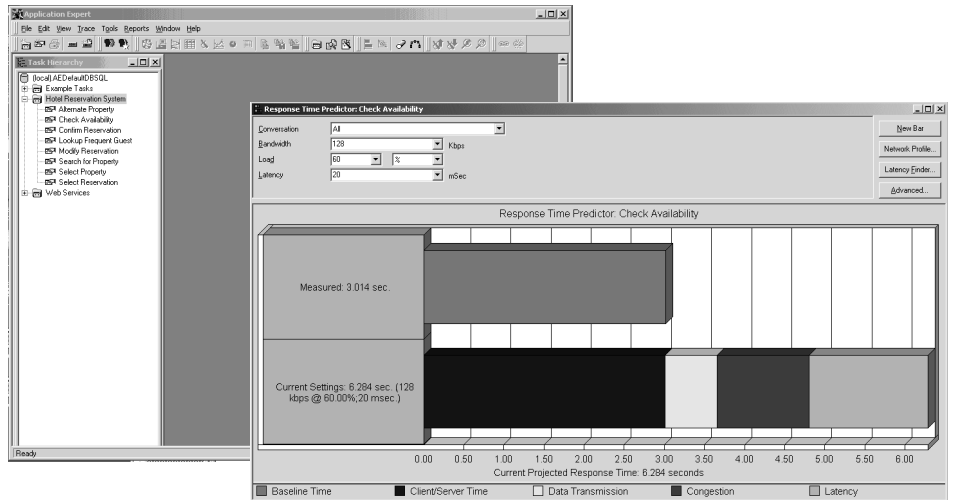


Figure 5: Transaction-level WAN provisioning. Application Expert analyzes how WAN bandwidth, latency and load affects end-user response time service levels.

It is helpful to think of WAN links not simply in terms of bandwidth but rather in terms of the quality of service they provide. While some applications perform proportionally to the bandwidth they receive, others may be more sensitive to latency. WAN provisioning provides the information needed to define the specific QOS policies (such as priority queuing) based on the needs of the application.

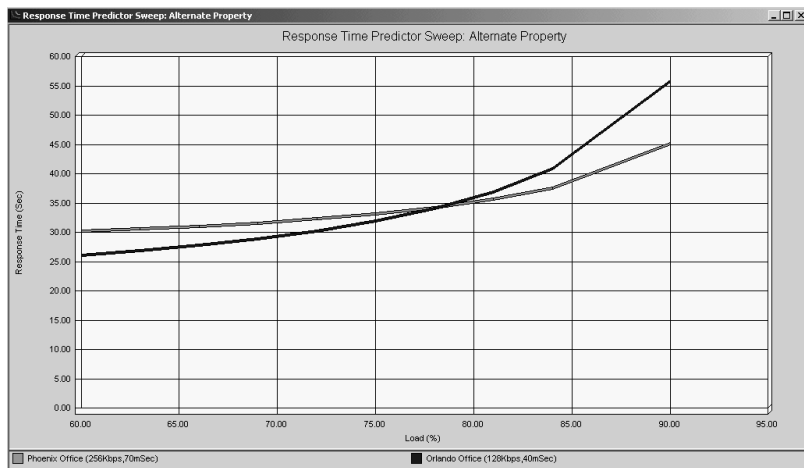


Figure 6: Application Expert's RTP Sweep feature helps determine the network quality of service (QOS) required to meet given response time requirements.

The location level—For multiple concurrent users of an application at a single site, or for a number of remote sites that connect directly to the application server (without link aggregation), WAN provisioning will include more sophisticated user profiles. Using Application Expert’s WAN provisioning module, recommendations for bandwidth are evaluated in terms of both end-user response time and projected link utilization. Because we are evaluating bandwidth requirements at this level for a single location, complex topology and extensive background traffic descriptions are not necessary. The WAN provisioning module provides a set of pre-defined network paths, or topologies, that can support any location-level WAN capacity exercise. To perform WAN provisioning, the user is guided through the definition of the profiles without having to leave the familiar Application Expert work environment.

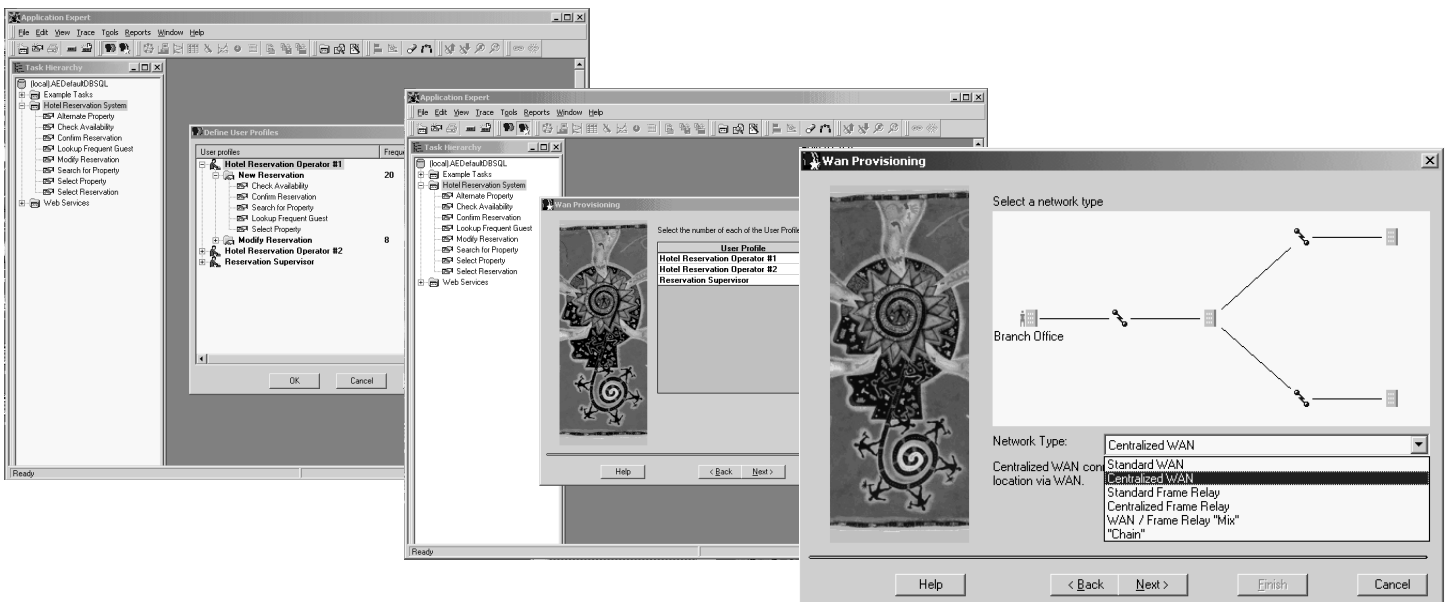


Figure 7: Location-level WAN provisioning. Application Expert user profiles are deployed onto easily customized infrastructure profiles using the WAN provisioning module.

For the location evaluated, the bandwidth report recommends bandwidth based on the user-defined target load, which may simply be a “best practice” standard (such as 60%), but could also be based on meeting end-user response time service levels. The response time report lists the transactions from the User Profile and identifies the projected response time and sensitivity of each. Since the reports are interactive, the user can immediately see how response time might change for different bandwidths by simply changing the bandwidth value in the table. Information gained by drilling down into Application Expert’s RTP and Thread Analysis features can also be used to analyze how individual transaction threads affect performance (for more information on the value of Thread Analysis, see the “Predictive Tuning for Oracle Applications” white paper available from your Compuware representative).

The enterprise level—For multiple users at multiple locations where more complex network topologies result in link aggregation, or for sophisticated “What-if?” questions such as alternative server locations or network topologies, WAN provisioning will include a more robust network topology definition. Leveraging the ease-of-use and rapid calculation strengths of Compuware’s Predictor, bandwidth recommendations are reviewed as to their impact on overall network performance, including trunk and backbone configurations. Different network topologies, optional server locations and projected traffic trends can be quickly analyzed. And, of course, the business metric of end-user response time may still be used as the primary decision metric.

Large-scale WAN provisioning must keep a sharp focus on the goals of WAN bandwidth analysis for new application deployment. It is important to note that this does not require the complexity of traditional, “general purpose” network modeling as discussed earlier, since this level of complexity would result in a bottleneck in the deployment process without adding any real value. In keeping with these goals, Predictor integrates with the Application Expert database for creating the user profiles, one of the key components of an accurate and successful analysis.

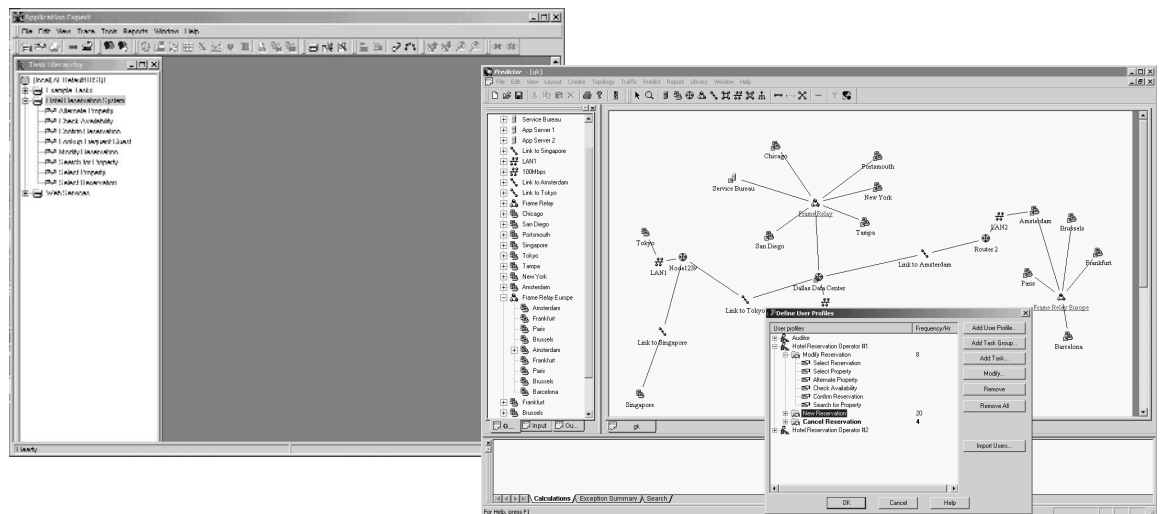


Figure 8: Enterprise-level WAN provisioning. User profiles are deployed into a WAN-centric topology model using Predictor.

Most enterprise application deployments can be abstracted to some degree to simplify this level of WAN provisioning. Phased application deployments can be analyzed one phase at a time; multiple sites with similar configurations can be analyzed as just one site; logical groups of sites can be “collapsed” into a single location. Keeping in mind the WAN focus will facilitate this approach and help you avoid the common pitfalls of modeling.

Conclusion

While there may be many terms for the process, a rapid application deployment solution is an accepted and implemented best practice for many companies whose competitiveness relies on the applications that run the business. Compuware's Application Expert is the de facto standard for application profiling; in fact, it's the only true profiling solution on the market.

With the release of Vantage 8.0, Compuware's comprehensive performance management solution for distributed applications, Compuware's WAN provisioning solutions become the only focused capacity planning solution integrated with rapid application deployment, enabling the proven cost savings associated with just-in-time network design.

Compuware products and professional services—delivering quality applications

Compuware is a leading global provider of software products and professional services which IT organizations use to develop, integrate, test and manage the performance of the applications that drive their businesses. Our software products help optimize every step in the application life cycle—from defining requirements to supporting production service levels—for web, distributed and mainframe platforms. Our services professionals work at customer sites around the world, sharing their real-world perspective and experience to deliver an integrated, reliable solution.

Please contact us to learn more about how our comprehensive products and services can help your organization improve productivity, create higher quality applications and ensure performance in production.

All Compuware products and services listed within are trademarks or registered trademarks of Compuware Corporation. All other company or product names are trademarks of their respective owners.
© 2002 Compuware Corporation



www.compuware.com