Submitted to IEEE OPENARCH'98, San Francisco, CA, April 1998

ANTS: A Toolkit for Building and Dynamically Deploying Network Protocols

David J. Wetherall, John V. Guttag and David L. Tennenhouse*

Software Devices and Systems Group
Laboratory for Computer Science
Massachusetts Institute of Technology

Abstract

We present a novel approach to building and deploying network protocols. The approach is based on mobile code, demand loading, and caching techniques. The architecture of our system allows new protocols to be dynamically deployed at both routers and end systems, without the need for coordination and without unwanted interaction between co-existing protocols.

In this paper, we describe our architecture and its realization in a prototype implementation. To demonstrate how to exploit our architecture, we present two simple protocols that operate within our prototype to introduce multicast and mobility services into a network that initially lacks them.

1 Introduction

The performance of modern distributed computing is increasingly dependent upon the network protocols used to move information among machines. Curiously, however, the evolution of these protocols has been much slower than than the evolution of almost any other part of the environment on which computing systems are built.

The slow evolution is attributable neither to lack of need nor to a lack of innovative ideas. In the case of IP, for example, changes are underway to better support multimedia applications, as well as to accommodate a larger number of potentially mobile hosts [6, 4, 5, 15]. Unfortunately, though agreement on the need for these changes was reached many years ago, they are still not fully deployed.

The problem is that the current process of changing network protocols is both lengthy and difficult. It requires standardization, since internetworking protocols are the basis of interoperability. This means that years may elapse between the time the need becomes apparent and the time consensus is reached on how to address that need. Furthermore, once the new protocol has been accepted, deployment is difficult. It must be done manually and in a backwards compatible fashion, since there is no automatic mechanism for upgrading functionality or dealing automatically with multiple protocols.

This paper presents a new approach to network protocol innovation that addresses all of these problems. The essence of our approach is to standardize a communication model (rather than individual communication protocols) that allows uncoordinated deployment of co-existing protocols. We have developed an active network [18, 19] toolkit, called ANTS, in which new protocols are automatically deployed at both intermediate nodes and end systems by using mobile code techniques. Our architecture views the network as a distributed programming system by providing a programming language model for expressing new protocols in terms of operations at nodes. Compared with alternative systems in which new protocols may be formed by selecting from a library of components, e.g., the x-kernel [11], ANTS provides the greater flexibility that accompanies a programming language and the convenience of dynamic deployment.

In the next section of this paper, we present the ANTS protocol architecture. We then demonstrate how the architecture can be exploited by presenting simple protocols that support multicast and mobility, two directions in which IP is currently being extended. This is followed by a discussion of our prototype implementation of the the ANTS architecture and some comments about a new implementation to be distributed in August of 1997. We then contrast our system with related work, and offer conclusions and suggestions for further work.

*djw@lcs.mit.edu. http://www.sds.lcs.mit.edu/. This work was supported by DARPA, monitored by the Office of Naval Research under contract No. N66001-96-C-8522, and by seed funding from Sun Microsystems Inc.

2 ANTS Protocol Architecture

An ANTS-based network consists of an interconnected group of nodes that execute the ANTS runtime; the nodes may be connected across the local or wide area and by point-to-point or shared medium channels. The system builds on the link layer services of the channels to provide network layer services to distributed applications.

Unlike IP, the network service provided by ANTS is not fixed – it is flexible. Different applications are able to introduce new protocols into the network by specifying the routines to be executed at network nodes that forward their messages. Thus ANTS is a distributed programming system as much as it is a network service. Applications may customize network processing to suit their needs by pushing processing into the network – either processing that is traditionally performed at end-systems or novel kinds of processing that only make sense in the context of active networks.

In designing ANTS, we set three goals for network protocol innovation. All describe more flexible forms of innovation than are currently achieved in the Internet.

- The nodes of the network must simultaneously support a variety of different network protocols.
- The architecture must support the construction of new protocols by mutual agreement among interested parties, rather than requiring new protocols to be registered in a centralized manner. We do not expect all users to construct new protocols directly, but rather to choose between protocols offered by third party software vendors.
- The architecture must support the dynamic deployment of new protocols, since it is unreasonable to take portions of the network "off-line" in order to configure nodes to support new protocols– especially as the scale of the network increases.

Our architecture meets these goals through the use of three key components.

- The packets found in traditional networks are replaced by *capsules* that refer to the processing to be performed on their behalf.
- Routers and end nodes are replaced by *active nodes* that execute capsules and maintain state.
- A *code distribution* mechanism ensures that processing routines are automatically and dynamically transfered to those nodes where they are needed.

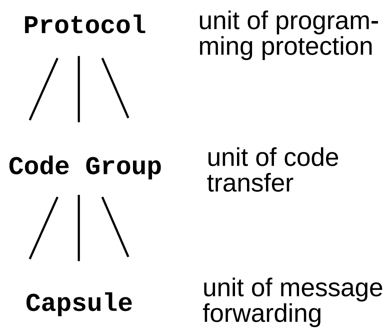

Figure 1: Capsule Composition Hierarchy

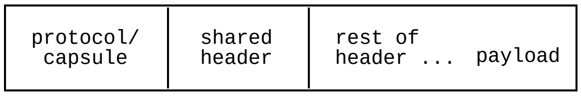

Figure 2: Capsule Format

2.1 Protocols and Capsules

To make use of programmable network elements, we require a model for combining forwarding routines at individual nodes into a pattern of behavior – a protocol – that defines the processing to occur across the network as a whole. Further, the model must separate patterns of behavior from each other.

In ANTS, we do this using capsules, code groups, and protocols. The relationships between these entities is illustrated in Figure 1.

- A *capsule* is a generalized replacement for a packet. Its most important architectural function is to include a reference to the forwarding routine to be used to process the capsule at each active node. Some forwarding routines are "well-known" in that they are guaranteed to be available at every active node. These primarily include routines for common case processing, i.e., unreliable data transfer with standard routing, and for bootstrapping network services, such as the code distribution scheme to be described shortly. Other routines are "application-specific". Typically, they will not reside at every node, but must be transfered to a node by the code distribution scheme before capsules of that type can be processed.
- A *code group* is a collection of related capsule types whose forwarding routines are transferred as a unit by the code distribution system.
- A *protocol* is a collection of related code groups that are treated as a single unit of protection by the ac-

tive nodes. Thus protocols are the units by which the network as a whole is customized by applications, and capsules within a single protocol will typically manipulate shared information.

Capsule Format

The format of capsules as they are carried across link-layer channels is sketched in Figure 2. Each capsule carries an identifier for its protocol and particular capsule type within that protocol. The identifier is based on a fingerprint (e.g., the MD5 message digest) of the protocol code. It is used for demultiplexing to a forwarding routine in the same sense as the Ethernet type and IP version and protocol fields.

That the capsule identifier is derived from the code description of the protocol of which it is a part is crucial for two reasons:

- It greatly reduces the danger of protocol spoofing. This is because a fingerprint based on a secure hash is effectively a one-way function from code to identifier. Each active node can verify for itself (and without trusting external parties) that a particular set of programs maps to a given identifier. For this reason, our architecture uses protocols as the unit of protection, preventing one protocol from interfering with the state of another within active nodes.
- It allows protocols and capsule types to be allocated quickly and in a decentralized fashion, since their identifier depends only on a fingerprint of the protocol code.

The remainder of the capsule format is comprised of a shared header that contains fields common to all capsules, a type-dependent header that may be updated as the capsule traverses the network, and a payload. The important components of the shared header are source and destination addresses and information about resource limits to be enforced by nodes. We use IPv4 style addresses for convenience. The remainder of each capsule varies according to its type.

2.2 Active Nodes

A key difficulty in designing a programmable network is to allow nodes to execute user-defined programs while preventing unwanted interactions. Not only must the network protect itself from runaway protocols, but it must offer co-existing protocols a consistent view of the network and allocate resources between them.

Our approach has been to execute protocols within a restricted environment that limits their access to shared resources. Active nodes play this role in our architecture. They export a set of primitives for use by application-defined processing routines, which combine these primitives using the control constructs of a programming language. They also supply the resources shared between protocols and enforce constraints on how these resources may be used as protocols are executed. We describe our node design along these two lines.

Node Primitives

We chose an initial set of primitives based on our experience with a predecessor system [22]. This work suggests that a relatively small set of primitives is sufficient to express a number of different and useful forwarding routines. We support the following categories of node primitives:

- *environment access*, to query the state of links and routing tables at the local node;
- *capsule manipulation*, with access to both header fields and payload;
- *control operations*, to allow capsules to forward, discard, and suspend/wake themselves;
- *caching*, to store and access soft-state, i.e., objects with application-defined meanings that are held for a short interval; and
- *rendezvous*, to allow multiple capsules to meet each other and coordinate their actions.

The set of primitives available at active nodes is important because it determines the kinds of processing routines that can be composed by applications. For example, without the ability to store and access node state, individual capsule programs will be unable to communicate with each other. Further, the compactness and execution efficiency of capsule programs is affected by these primitives. Both are enhanced if the primitives are a good match for the processing, and degraded otherwise. For example, the neighbors at a given node may be found either by walking the entire routing table looking for adjacent nodes, or by asking the question directly of the node, depending on which topological queries are supported. The direct query can be represented compactly and executed efficiently as a built-in node primitive, while the other program cannot.

Execution Model

Given the preceding primitives, the node must provide an execution model that supports them while meeting network security and resource management goals.

When a capsule arrives at a node, its associated processing routine is run to completion (unless it exceeds its resource limit). The routine processes the payload of the capsule and initiates any further actions, e.g., forwarding, that are necessary. Unlike more general mobile agent systems, the node provides no support for migrating computations at arbitrary points during execution. Instead, processing routines may update capsule fields and store application-defined information in the shared node cache of soft-state. Together, these mechanisms allow the construction of computations that evolve their behavior as they traverse the network.

During capsule processing, active nodes are responsible for the integrity of the network and handle any errors that arise. Since capsule processing resembles a distributed programming system in which there are many legitimate users with small tasks, authentication and other traditional security schemes are likely to be too heavyweight to be used for common-case forwarding programs. Instead, we rely on the safety mechanisms of mobile code technologies (e.g., sandboxing and Java bytecode verification) to execute untrusted routines efficiently in a contained manner. Conversely, the occasional use of primitives that manipulate shared logical resources, e.g., updates to the default routing tables, must be authenticated.

This model is not yet sufficient, however, to ensure that the network is robust, nor that its resources are allocated in an intended manner. For these purposes, additional mechanisms are needed to limit the physical resources consumed by capsule programs, both at individual nodes and across many nodes.

We associate with each capsule a resource limit that functions as a generalized TTL (Time-To-Live) field. This limit is carried with the capsule and decremented by nodes as resources are consumed; only nodes may alter this field, and nodes discard capsules when their limit reaches zero. Unlike the TTL field used in IP today, the resource limit is divided across capsules when one capsule creates others within the network. This is required to reason about total resource bounds.

At individual nodes, it is straightforward to charge for resources as they are consumed. Processing time and link bandwidth are allocated by time and capsule quanta, respectively; node memory is allocated by cached objects, since caching converts memory into a renewable

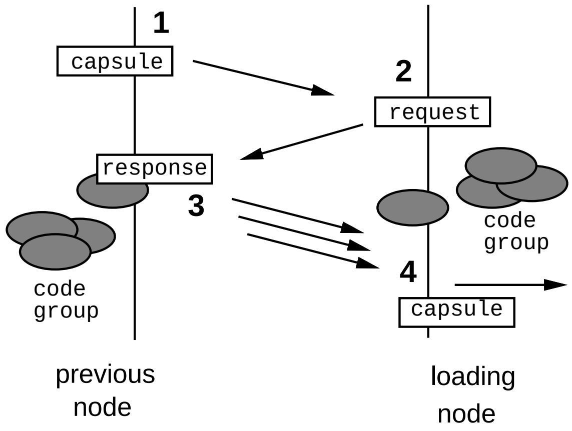

Figure 3: Demand Loading of Code Groups

resource. We hope, however, that it will prove feasible to enforce static limits at nodes with a scheme similar to [7] or by using proof-carrying code techniques [13].

2.3 Code Distribution

The third component of our architecture is a code distribution system. Given a programmable infrastructure, a mechanism is needed for propagating program definitions to where they are needed. A good scheme must be efficient, adapt to changes in node connectivity, and limit its activity so that the network remains robust.

Many different mechanisms are possible. At one extreme, programs may be carried within every capsule. This scheme is only suited to transferring extremely short programs when bandwidth is not at a premium. At the other extreme, programs may be pre-loaded into all nodes that may require them by using an out-of-band or management channel prior to using a new protocol. This scheme is not suited to our goals of rapid and decentralized deployment.

Instead, our approach has been to couple the transfer of code with the transfer of data as an in-band function. We believe this has several advantages. It limits the distribution of code to where it is needed, while adapting to node and connectivity failures. It improves startup performance and facilitates short-lived protocols by overlapping code distribution with its execution. It further suits our research goals by allowing customized processing to be expressed at a fine granularity, i.e., per capsule rather than per application session.

We have designed a scheme that is suited to flows, i.e., sequences of capsules that follow the same path and require the same processing. By analogy with traditional networks, we believe that this is a reasonable common case. It is motivated by two observations: networked applications typically exchange many related messages; and the peer relationships of layered protocol models result in related packet processing at different locations.

At end-systems, applications may begin to use a new protocol at any time by registering the code definition at their local node. Capsules of the new type may then be injected into the network and received from it. As capsules travel through nodes of the network, a lightweight protocol is used to transfer the capsule programs incrementally from one node to the next. For this purpose, capsules must be organized into code groups according to their dependencies. If one type of capsule refers to another type, their definitions are grouped for joint transfer.

A sequence of events that illustrates the operation of this lightweight protocol is listed below and shown in Figure 3.

1. Capsules identify their type and the protocol to which they belong as they travel. This information is immutable for a given instance of a capsule.
2. When a capsule arrives at a node, a cache of protocol code is checked. If the required code is not present, a load request based on the capsule type and protocol is sent to the "previous" node, i.e., the node from which the capsule arrived. The capsule execution is suspended, awaiting the code, for a finite time.
3. When a node receives a load request that it can answer, it does so immediately. It sends load responses that contain the entire code group that is implicated.
4. When a node receives a load response, it incorporates the code into its cache. If the code group is fully loaded, it wakes sleeping capsules. If the required responses are not forthcoming, sleeping capsules are discarded without further action.

This scheme has some important properties. First, the reliance of a node on the "previous" node is designed to draw code from a source node along the network paths where it is needed. As many capsules are transferred, a region is grown, within which the same processing is invoked repeatedly and code transfer is no longer necessary. If network paths change, then code transfer will resume in order to adapt to the new connectivity.

Second, the connectionless nature of the scheme is designed to provide rapid loading without concern for reliability. If the size of a code group is bounded, then the amount of work that the network will perform in response to capsules of unknown type is bounded to within a constant factor. The intent is to ensure that the network will remain robust under high load, and it is the principal reason that we do not use higher-level connections, e.g., TCP. In effect, this scheme reflects the restrictions of operating within the network layer. Capsules will be discarded when load requests or responses are lost or too slow, as well as in response to congestion. In this case, higher-level processing at the end-systems must intervene to provide the required reliability, just as it does today.

We believe that our code distribution scheme has qualities that will prove it efficient, adaptive, and robust, but this must be borne out by experimentation. In order for it to best accommodate the largest number of scenarios, we include two special cases. First, for very small code groups the code may be carried along with every capsule if desired. Second, capsules may be constructed to "prime" a path with a code group to improve the startup period.

3 Programming with ANTS

To demonstrate how we intend our architecture to be used, we describe two simple protocols that introduce multicast and mobility services into an ANTS network that initially lacks them. We chose these examples because they represent two areas of widespread interest in which the Internet community is currently dealing with the difficulties of innovating protocols.

The presented protocols were written and tested using the Java-based prototype implementation of ANTS described in Section 4. In developing the protocols, it was not our intent to present new and better solutions to these particular problems. Our goal was merely to demonstrate how our approach may be used to write protocols that address these kinds of problems in a number of different ways, depending on application requirements rather than relying on a "one size fits all" solution.

3.1 Mobile Hosts

We introduce support for mobile hosts into an ANTS network with a `Mobile` protocol composed of two cooperating capsule types. One type of capsule is sent by the mobile host to register forwarding information while it is roaming. The second is used by other hosts to send messages to the mobile host. To be consistent with mobility schemes, we use the notion of home and foreign agents. The home agent is used to intercept messages at

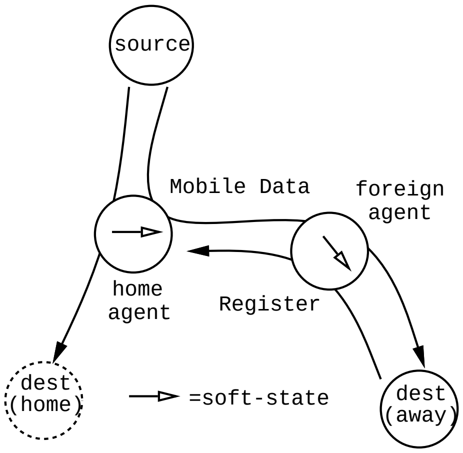

Figure 4: Mobile Capsule Paths

```
// on entry:
//   home = home agent
//   next = node at which to register
//   forward = address to be registered

// go to foreign and then home agent
if (n.address() != next) {
  n.routefornode(this, next);
  return;
}

// insert a forwarding address
n.put(src, MOBILE, new W_N(forward), IDLE);

// after doing foreign, do home
if (n.address() != home) {
  forward = next; next = home;
  n.routefornode(this, next);
}
```

Figure 5: Mobile `Register` Capsule

the base location of the mobile host. The foreign agent is used as a "care of" address to reach the mobile host while it is away from its base. The paths of these two types of capsule is shown in Figure 4 and their code in Figures 5 and 6.

Mobile hosts that are roaming periodically send `Register` capsules to their home agent via a local foreign agent. The program carried by this capsule updates forwarding addresses cached at the home and foreign agents. In each, an updated forwarding pointer is entered into the node cache; the home agent forwards to the foreign agent, and the foreign agent to the current mobile location. At the home agent, `Register` capsules are silently discarded, having established their forwarding pointers within the network. As the mobile moves, old forwarding pointers will either be supplanted by fresh information or evicted from the cache after a brief interval.

```
// look up forwarding record
W_N f = (W_N)n.get(dst, MOBILE);

// if found, update our route
if (f != null) next = f.node;

// and continue on our way
if (n.address() != next)
  n.routefornode(this, next);
if (n.address() == dst)
  n.delivertoapp(this, dpt);
```

Figure 6: `Mobile Data` Capsule

To communicate with the mobile host, other hosts send `Mobile Data` capsules that make use of this forwarding information. This capsule program is directed by default routing towards the base location of the mobile. If the mobile is at home, the capsule will reach it and be delivered. If the mobile is roaming, the capsule will discover a forwarding pointer as it traverses the home agent, and follow it to the foreign agent. There, it will find a further pointer to the current mobile location and so be delivered.

Despite the simplicity of this scheme, it provides the essential feature of mobility: hosts may be reached as they move without introducing another layer of addressing. There are also some interesting comparisons with Mobile IP [15]. First, unlike Mobile IP, a different protocol is used to reach mobile and non-mobile hosts. This is the nature of our approach–to introduce and use specialized protocols. To assist applications in selecting which protocol to use in the first place, a directory service may be used, much as the choice of IPv4 versus IPv6 is to be incorporated into the DNS [9]. Second, it is not necessary to confine mobile forwarding information to the edges of the network. To facilitate shortcut routing, mobile updates may enter forwarding pointers at any node, and messages from other nodes will follow them like a trail of crumbs once their paths cross.

3.2 Multicast

We introduce a basic `Multicast` protocol, resembling IP multicast [6], composed of two cooperating capsule types[1]. One type of capsule is sent to subscribe to a group, and the other carries the multicast message itself.

Applications that wish to receive messages sent to a given group by a particular sender periodically send `Subscribe` capsules towards the sender. The program carried by this capsule installs (or refreshes) forwarding

[1] Readers familiar with IP multicast will note that our multicast provides a somewhat different service. This is discussed after the scheme is presented.

```
// on entry:
//   group = multicast group
//   sender = multicast sender
//   reverse = last visited node

// look up forwarding record
W_JAN m = (W_JAN)n.get(group,sender,MCAST);

// or make a new one if necessary
if (m == null) {
  m = new W_JAN();
  n.put(group, sender, MCAST, m, IDLE);
}

// are we at an intermediate node?
add: if (reverse != 0) {
  if (m.nodes == null) {

    // start a new list
    m.nodes = new Node[1];
    m.nodes[0] = reverse;
  } else {

    // does it contain our info?
    for (int i = 0; i < m.nodes.length; i++)
      if (m.nodes[i] == reverse) break add;

    // if not, add it
    int len =  m.nodes.length;
    Node[] nn = new Node[len+1];
    System.arraycopy(m.nodes,0,nn,0,len);
    nn[len] = reverse; m.nodes = nn;
  }
}

// need to refresh upstream entry?
long time = System.currentTimeMillis();
if (time - m.time < RATE) return;
m.time = time;

// if so, update route and continue
if (n.address() != sender) {
  reverse = n.address();
  n.routefornode(this, sender);
}
```

Figure 7: Multicast `Subscribe` Capsule

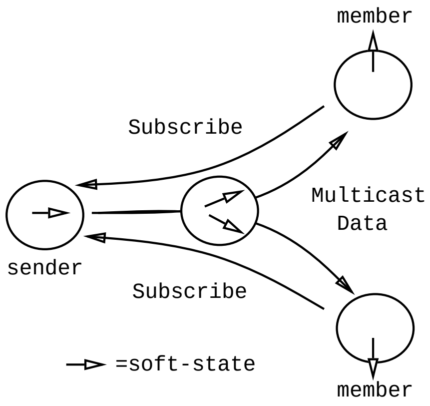

Figure 8: Multicast Capsule Paths

```
// look up forwarding record
W_JAN m = (W_JAN)n.get(group,sender,MCAST);

// must find it to continue
if (m != null) {
  if (m.nodes != null) {

    // send a copy every way
    for (int i = 0; i < m.nodes.length; i++)
      n.routefornode(this, m.nodes[i]);
  } else

    // or deliver to application
    n.delivertoapp(this, dpt);
}
```

Figure 9: `Multicast Data` Capsule

pointers that are cached in each router it traverses; forwarding information sent by different nodes is merged to form a distribution tree. To multicast to the group, the sender node sends a `Multicast Data` capsule that simply routes itself along the distribution tree. The paths of these capsules are shown in Figure 8 and the code in Figures 7 and 9.

The `Subscribe` program begins by looking up the forwarding record for the group in the node cache, creating a fresh record if none is found. To separate this forwarding record from other information in the cache (including other multicast sessions), the record is stored under a key that is the combination of the group address, sender address and multicast capsule type. Once the forwarding record is located, a "reverse" pointer in the direction of the subscriber is merged into the forwarding record. At the leaf subscriber node itself, the forwarding record is created with an empty forwarding list.

The forwarding method of the `Multicast Data` capsule makes use of the forwarding records it finds at nodes, sending a copy of itself along every indicated "reverse" path found in the record at each node. If no forwarding information can be found, the capsule is discarded. At end-systems, where there is an empty forwarding record, the capsule delivers itself to an application.

Together, these capsule programs implement an unreliable multicast protocol with the central property of network-based multicast: efficient use of bandwidth. The service also differs from IP multicast in two interesting respects. First, the scheme is localized to the nodes using the protocol, and does not require that multicast-capable routers be separately identified or organize themselves into a tree. Second, it provides a different multicast primitive, since members subscribe to the combination of a group and sender. If multiple senders are needed, then multiple distribution trees may be formed by having members subscribe to each of the senders. Alternatively, the sender may be considered the root of a core-based tree [2], with messages routed up the tree towards the root and down other branches.

4 Prototype Implementation

Over the last several months we have been experimenting with a prototype implementation of the ANTS architecture. The implementation was designed primarily to allow us to evaluate the suitability of our approach to creating and deploying protocols. We have used it to test and debug the `Mobile` and `Multicast` protocols discussed in this paper, as well as a high performance reliable multicast developed by Li Lehman and a TCP SYN-flooding defense protocol developed by Van Van [21].

We are now in the process of building a new implementation. The structure of the new implementation is quite similar to the one described here, but far more attention is being devoted to performance related issues. Also, security and resource management, which were largely ignored in our prototype, are being addressed.

The current implementation is written in Java and runs as a user-level process under Linux. The code distribution protocol transfers processing routines in Java class-file format. We chose Java because of its support for safety and mobility (through bytecodes and their verification) and the likely emergence of higher performance compilers and runtimes. Its flexibility as a high-level language and support of dynamic linking/loading, multi-threading, and standard libraries has allowed us to evolve our design while maintaining a small code base (≈5000 lines).

The major components of the ANTS architecture are implemented using the classes listed in Table 1.

Class	Key Methods
`Node`	`address`, `get`, `put`, `routefornode`, `delivertoapp`
`Channel`	`send`, `receive`, `node`
`Application`	`send`, `receive` (upcall), `node`
`Capsule`	`evaluate`, `length`, `serialize`, `deserialize`

Table 1: Key Classes and Methods

When an ANTS node is started, its root thread instantiates a single `Node` object and one `Channel` object for each network interface. The node creates one thread for cache maintenance and other management tasks.

Though we prefer to think of each protocol running in its own address space with one thread per capsule execution, this is not what actually happens. Creating a thread for each capsule would impose too high a performance penalty, even for a prototype. Instead, we assume that most capsule processing routines are short and execute them sequentially using one thread per channel. If the operation is long-lived, then we lazily spawn additional threads.

When a packet arrives from the link layer, the channel thread attempts to convert it to a capsule instance of the appropriate class. If the required code is not present at the node, then the packet is retained in the node cache while the code is fetched using the code distribution protocol. Once a capsule instance is created, the thread calls its `evaluate` method, passing the node instance as a parameter. As it is evaluated, the capsule code has access to the private soft state associated with that instance of the protocol as well as the public state (e.g., routing tables) of the node.

`Node` Class

The `Node` class represents the runtime of a single network node, including its caches and code distribution protocol. It provides a set node primitives that can be invoked by capsule programs. These primitives allow access to the state at the node and enforce various security constraints.

Table 1 lists some key methods, including `routefornode`, which forwards a copy of a capsule towards a given destination, and `get` and `put`, which are used to manipulate that part of the node cache that can be directly accessed by protocols. All the node caches are managed in a least-recently-entered order. This method was chosen because the primary goal of this implementation was to help us understand how to write correct protocols, and this cache management strategy puts more stress on protocols than the better performing least-recently-used strategy. We also implemented a timeout in cache entries. Not only does this prevent the network from maintaining stale state (e.g., old session ids), but by shortening the time out we are again able to put stress on protocols.

`Channel` Class

The `Channel` class provides the interface to the link layer, connecting nodes via point-to-point or shared medium channels. At present, either Ethernet or UDP "tunnels" may be used to transfer capsules. These choices allow small networks to be constructed by running one node per machine and connecting the nodes with Ethernet channels. Larger networks are emulated by running many nodes per machine and connecting the nodes with UDP channels.

`Capsule` Class

The `Capsule` class is a virtual class that can be specialized to create the capsule types that comprise protocols. During capsule processing at nodes, each packet received from the link layer is manipulated as an instance of its corresponding `Capsule` subclass. In our prototype implementation, if an error occurs, execution of the capsule is terminated and the state associated with that execution of the capsule is released. It would be straightforward to extend this recovery process with an error message scheme analogous to ICMP.

In addition to providing the base class for new protocols, our current implementation provides several built-in subclasses. The class `DataCapsule` allows applications to transfer data using default (i.e., shortest path) routes. The system classes `DLRequestCapsule` and `DLResponseCapsule` are used by the code distribution protocol. They provide the bootstrapping capability needed to install other protocols.

`Application` Class

Programs that use the ANTS service are constructed by specializing the `Application` class. This is a container for end-system processing that provides a small API for registering capsules, injecting capsules into the network and receiving capsules from the network. It runs within the same address space as the node to which it is directly attached. At nodes internal to the network, it can be used to implement SNMP-like node management applications. At end-systems, it provides a bridge to the end user.

Measurement and Evaluation

Though our prototype implementation was not built with performance in mind, we did run a small number of performance tests. The goal of these was to gain some insight into the performance impact of various architectural decisions.

Table 2 shows the latency and throughput across a single node. The tests were run on Pentium Pros (200MHz) running Linux and connected by 100 Mbps Ethernet. The parameters of the node cache were adjusted so that demand loading was infrequent. The row labeled `Data` shows results for a minimal capsule program using built-in routing. It is our equivalent of a "null RPC" and represents the baseline performance of our prototype. The second and third rows show results for the example protocols presented in Section 3.

Capsule	Latency (ms)	Throughput (capsules/sec)
Data	3.1	160
Mobile	3.2	150
Multicast	4.1	130 (x2)

Table 2: Node Processing Measurements

Capsule	Size (bytes)	Latency w/ Load (ms)
Data	538	13.1
Mobile	1625	17.5
Multicast	1643	18.4

Table 3: Code Distribution Measurements

As expected, the absolute baseline performance was not good. We pay some penalty (an average of 300 to 600 microseconds depending on packet length) for operating at user level. We pay much larger penalties for an inefficient implementation of our runtime environment and for using an interpreter to execute capsule code. Experience with another implementation of the ANTS architecture [12] demonstrates that a more careful implementation of the runtime environment (especially with respect to memory management) can increase throughput by a factor of at least four. Using a better implementation of the Java runtime may make an even bigger difference, but we don't yet have any experience to cite.

Table 3 shows the cost of distributing code in terms of the size of capsule programs and the latency of processing capsules when their programs must be demand loaded. Though the size of capsule programs is considerably larger than is necessary because our implementation uses the Java classfile format directly, the example routines are short. This suggests that realistic processing routines can be transferred without consuming much bandwidth. Additionally, the latency attributable to demand loading seems quite reasonable. Even in our untuned prototype implementation, the latency attributable to delivering and loading the capsule code is comparable to the latency associated with establishing a connection on conventional networks.

5 Related Work

We believe our approach is novel in its application of mobile code, demand loading, and caching techniques to the network layer.

The most similar recent work we are aware of is the messenger paradigm [8] and work on flexible protocol stacks that preceded it [20]. Like our system, this work allows new protocols to be deployed. The intent, however, is to investigate the structuring of communicating systems, including distributed operating systems and intelligent agents. As such, it lacks the network layer specializations, e.g., demand loading, that we have developed.

Some modern protocol architectures have been configurable, as opposed to programmable. The x-kernel [11] provides a collection of micro-protocols from which protocols (e.g., RPC) can be synthesized. Configurable systems can further increase their flexibility by deferring the selection of components until runtime, and so the x-kernel supports the dynamic composition of micro-protocols on a per packet basis. Although configurable systems are capable of expressing a range of protocols, their means of composition, e.g., layering, is less flexible than that of a programming language.

The earliest programmable network based on mobile code that we are aware of is Softnet [24], an experimental packet radio network constructed in the early 1980s. Its goal was similar to our own: to allow users to define their own high level services. As with our approach, packets were considered to be programs of a language, FORTH, and interpreted at nodes on arrival. Softnet is an intriguing example of a real programmable network that inspired a user community and workshops, but unfortunately fell into disuse with little documented about its successes and failures. We speculate that this was because of difficulties with safety and efficiency, problems that may now be more tractable, given the recent advances in mobile code and operating system technology.

End-to-end code shipping to improve performance has been studied in the context of RPC [17, 14]. Our approach offers a greater scope for customization by including intermediate nodes as well as end-systems.

There are several new efforts investigating active networks. The Switchware project [16] is developing a programmable switch approach to explore the use of formal methods to assure network security. The Active Bridge [1] is their first example of such a programmable network element. The Netscript project [23] is addressing network management tasks (such as routing, packet analysis, and signaling) with a dataflow language and scripting agents that may be dynamically distributed. The Liquid Software project [10] is developing technologies (such as the ultra-fast compilation of Java bytecodes) that will facilitate the use of mobile code in the network substrate. An effort at Georgia Tech [3] is applying active networking concepts to congestion control and other areas. Finally, the Smart Packets project at BBN is exploring the use of packets that carry code for network configuration and control purposes.

6 Conclusions

In this paper, we presented an architecture that supports building and dynamically deploying network protocols. In contrast to a standardization-based process, our approach:

- Allows new protocols to be automatically, dynamically, and rapidly deployed to exactly those nodes in the network they are needed; and,
- Requires no advanced consensus about the kinds or definitions of the protocols.

We achieved these results by treating the network as a distributed programming system and through the application of mobile code, demand loading and caching techniques.

Our Java-based prototype allowed us to experiment with the ANTS network programming model and test its code distribution system. In addition to the simplistic examples presented here, we have used ANTS to implement several more complex protocols, including a high performance reliable multicast. With this experience, we are now building a new implementation, to be distributed in August of 1997, that pays greater attention to performance, security and resource management issues. Based on experience with a separate implementation [12], we expect that throughput can be increased by a factor of at least four by restructuring our Java code, and further still by using a better Java runtime, e.g., one including just-in-time compilation.

Our early experience with ANTS strongly suggests that this use of mobile code technologies has considerable promise. It can provide the means for automatically upgrading network protocols. This, in turn, can remove barriers to innovation, stimulate experimentation, and hasten the arrival of new functionality. It is interesting to speculate how many other protocols would be deployed were it not for the barriers to innovation that we are addressing.

Acknowledgments

We thank the members of our research group for much useful feedback as the design of our system progressed. In particular, we wish to acknowledge David Murphy

for his work on a separate node implementation and his contributions to an earlier version of this paper.

References

[1] D. S. Alexanger et al. Active Bridging. In *SIGCOMM'97*, 1997.

[2] A. Ballardie et al. Core based Trees. In *SIGCOMM'93*, 1993.

[3] S. Bhattacharjee et al. An Architecture for Active Networking. In *High Performance Networking (HPN'97)*, 1997.

[4] R. Braden et al. Resource ReSerVation Protocol (RSVP) – Version 1 Functional Specification. Internet Draft, Nov 1996.

[5] S. Deering and R. Hinden. Internet Protocol, Version 6 (IPv6) Specification. Request For Comments 1883, Dec 1995.

[6] S. E. Deering. Host Extensions for IP multicasting. Request For Comments 1112, Aug 1989.

[7] P. Deutsch and C. A. Grant. A Flexible Measurement Tool for Software Systems. In *Information Processing*, 1971.

[8] G. Di Marzo et al. The Messenger Paradigm and its Impact on Distributed Systems. In *Workshop on Intelligent Computer Communication*, 1995.

[9] R. Gilligan and E. Nordmark. Transition Mechanisms for IPv6 Hosts and Routers. Request For Comments 1933, April 1996.

[10] J. Hartman et al. Liquid Software: A New Paradigm for Networked Systems. Technical Report TR96-11, Dept. of Computer Science, Univ. of Arizona, 1996.

[11] N. C. Hutchinson and L. L. Peterson. The x-Kernel: An Architecture for Implementing Network Protocols. *IEEE Transactions on Software Engineering*, 17(1):64–76, Jan 1991.

[12] D. Murphy. Building an Active Node on the Internet. M.Eng Thesis, Massachusetts Institute of Technology, June 1997.

[13] G. Necula and P. Lee. Safe Kernel Extensions Without Run-Time Checking. In *2nd Symp. on Operating System Design and Implementation*, 1996.

[14] C. Partridge. *Late-Binding RPC: A Paradigm for Distributed Computation in a Gigabit Environment.* PhD thesis, Harvard University, 1992.

[15] C. Perkins, Ed. IP Mobility Support. Request For Comments 2002, Oct 1996.

[16] J. Smith et al. SwitchWare Accelerating Network Evolution. Technical Report MS-CIS-96-38, CIS Dept., Univ. of Pennsylvannia, May 1996.

[17] J. W. Stamos and D. K. Gifford. Remote Evaluation. *ACM Transactions on Programming Languages and Systems*, 12(4):537–565, Oct 1990.

[18] D. Tennenhouse et al. A Survey of Active Network Research. *IEEE Communications Magazine*, pages 80–86, Jan 1997.

[19] D. L. Tennenhouse and D. Wetherall. Towards an Active Network Architecture. In *Multimedia Computing and Networking 96*, 1996.

[20] C. Tschudin. Flexible Protocol Stacks. In *SIGCOMM'91*, 1991.

[21] V. C. Van. A Defense Against Address Spoofing Using Active Networks. M.Eng Thesis, Massachusetts Institute of Technology, June 1997.

[22] D. J. Wetherall and D. L. Tennenhouse. The ACTIVE IP Option. In *7th SIGOPS European Workshop*, 1996.

[23] Y. Yemini and S. da Silva. Towards Programmable Networks. In *FIP/IEEE Intl. Workshop on Distributed Systems Operations and Management*, 1996.

[24] J. Zander and R. Forchheimer. Softnet - An Approach to High-Level Packet Communication. In *ARRL 2nd Computer Networking Conference*, 1983.