P1 426568

THE UNITED STATES OF AMERICA

TO ALL TO WHOM THESE PRESENTS SHALL COME:

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office

July 11, 2001

THIS IS TO CERTIFY THAT ANNEXED HERETO IS A TRUE COPY FROM THE RECORDS OF THE UNITED STATES PATENT AND TRADEMARK OFFICE OF THOSE PAPERS OF THE BELOW IDENTIFIED PATENT APPLICATION THAT MET THE REQUIREMENTS TO BE GRANTED A FILING DATE UNDER 35 USC 111.

APPLICATION NUMBER: ***60/210,440***
FILING DATE: ***June 08, 2000***
PCT APPLICATION NUMBER: *PCT/US01/14760*

By Authority of the
COMMISSIONER OF PATENTS AND TRADEMARKS

L. Edelen

L. EDELEN
Certifying Officer

PRIORITY DOCUMENT
SUBMITTED OR TRANSMITTED IN COMPLIANCE WITH RULE 17.1(a) OR (b)

06-12-00 A/PROV

Express Mail mailing label no.EL212283154US
Date of Deposit: June 8, 2000
I hereby certify that this paper or fee is being deposited with the United States Postal Service "Express Mail Post Office to Addressee" service under 37 CFR 1.10 on the date indicated above and is addressed to the Assistant Commissioner for Patents, BOX PROVISIONAL APPLICATION, Washington, D. C. 20231.

By: _____
Karrie Martin

Attorney Docket No. AGLE0001PR

IN THE U.S. PATENT AND TRADEMARK OFFICE
Provisional Application Cover Sheet

Assistant Commissioner for Patents
BOX PROVISIONAL APPLICATION
Washington, D.C. 20231

Sir:

This is a request for filing a PROVISIONAL APPLICATION FOR PATENT under 37 CFR 1.53(b)(2).

INVENTOR(s)/APPLICANT(s)

Last Name	First Name	Middle Initial	Residence (City and Either State or Foreign Country
Calderone	Ted		San Carlos, California
Cook	Paul	M.	Woodside, California
Foster	Mark	J.	Palo Alto, California

Additional inventors are being named on separately numbered sheets attached hereto.

Title of the Invention

METHOD AND APARATUS FOR CENTRALIZED VOICE-DRIVEN NATURAL LANGUAGE PROCESSING IN MULTI-MEDIA & HIGH BAND WIDTH APPLICATIONS

Correspondence Address

Michael A. Glenn
3475 Edison Way, Ste. L
Menlo Park, CA 94025

Telephone No. (650) 474-8400

Enclosed Application Parts *(check all that apply)*

(X) Specification *Number of Pages* 24 (X) Small Entity Statement -(specify - IND or BUS)
and 3 Drawing(s)

( ) Other (specify)

Filing Fee and Method of Payment

X $75.00 for Small Entity _____ $150 for Large Entity

The Commissioner is authorized to charge the filing fee of $75.00 and any additional fees or credit any overpayment to Deposit Account No. 07-1445 (Order No. AGLE0001PR). A copy is enclosed for this purpose.

Respectfully Submitted,

EARLE. JENNINGS
Reg. No. 44,804

Attorney Docket No. AGLE0001PR

Applicant/Patentee: AGILETV CORPORATION
Serial or Patent No. _____ Atty Docket No. AGLE0001PR
Filed or Issued: Herewith
Title: METHOD & APARATUS FOR CENTRALIZED VOICE-DRIVEN NATURAL LANGUAGE PROCESSING IN MULTI-MEDIA & HIGH BAND WIDTH APPLICATIONS

VERIFIED STATEMENT (DECLARATION) CLAIMING SMALL ENTITY STATUS
37 CFR 1.9(f) and 1.27(c)--SMALL BUSINESS CONCERN

I hereby declare that I am
[ ] the owner of the small business concern identified below:
[X] an official empowered to act on behalf of the small business concern identified below:

NAME OF CONCERN: AGILETV CORPORATION

ADDRESS: 333 Ravenswood Ave., Bldg. 202, Menlo Park, CA 94025

I hereby declare that the above identified small business concern qualifies as a small business concern as defined in 13 CFR 121.3-18, and reproduced in 37 CFR 1.9(d), for purposes of paying reduced fees under 41(a) and (b) of Title 35, U.S. Code, in that the number of employees of the concern, including those of its affiliates, does not exceed 500 persons. For purposes of this statement, (1) the number of employees of the business concern is the average over the previous fiscal year of the concern of the persons employed on a full-time, part-time or temporary basis during each of the pay periods of the fiscal year, and (2) concerns are affiliates of each other when either, directly or indirectly, one concern controls or has the power to control the other, or a third party or parties controls or has the power to control both.

I hereby declare that rights under contract or law have been conveyed to and remain with the small business concern identified above with regard to the invention entitled: METHOD & APARATUS FOR CENTRALIZED VOICE-DRIVEN NATURAL LANGUAGE PROCESSING IN MULTI-MEDIA & HIGH BAND WIDTH APPLICATIONS, by inventor(s) Ted Calderone, Paul M. Cook, Mark J. Foster, described in

[X] the specification filed herewith.
[ ] application Serial No. _____ filed _____
[ ] patent # _____ issued _____

If the rights held by the above-identified small business concern are not exclusive, each individual, concern or organization having rights to the invention is listed below* and no rights to the invention are held by any person, other than the inventor, who could not qualify as a small business concern under 37 CFR 1.9(d) or by any concern which would not qualify as a small business concern under 37 CFR 1.9(d) or a nonprofit organization under 37 CFR 1.9(e).

*Note: separate verified statements are required from each named person, concern or organization having rights to the invention averring to their status as small entities. (37 CFR 1.27)

Name: _____
Address: _____
[ ] individual [ ] small business concern [ ] nonprofit organization

Attorney Docket No. AGLE0001PR

I acknowledge the duty to file, in this application or patent, notification of any change in status resulting in loss of entitlement to small entity status prior to paying, or at the time of paying, the earliest of the issue fee or any maintenance fee due after the date on which status as a small entity is no longer appropriate. (37 CFR 1.28(b)).

I hereby declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true; and further, that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under 1001 of Title 18 of the U.S. Code, and that such willful false statements may jeopardize the validity of the application, any patent issuing thereon, or any patent to which this verified statement is directed.

NAME OF PERSON SIGNING: James E. Jervis

TITLE IN ORGANIZATION: Vice President, Intellectual Property

ADDRESS OF PERSON SIGNING: 333 Ravenswood Ave., Bldg. 202 Menlo Park, Ca 94025

SIGNATURE James E. Jervis DATE June 8, 2000

AgileTv System

Current Practice

Currently voice operated functions using the latest voice recognition technologies have been limited to only a hand full of applications such as toys, appliances, some computers, voice dictation, cellular phones and voice control of one's home. Most of these applications use voice recognition technology that runs on a computer or voice recognition chip technology. These voice recognition systems typically offer only a limited number of commands and the recognition efficiency could be considered only fair and often requires voice training. There is however, another class of voice recognition technology called "Natural Language" which requires state of the art processing software and hundreds of megabytes of RAM to support. "Natural Language" voice recognition is currently being used in high end systems such as billing systems for the utility companies and the New York stock exchange because of it's ability to recognize spoken words from any voice. Some natural language systems claim to be totally user independent and are also capable of recognizing speech in several different languages. The system described in this disclosure uses a "Natural Language" voice recognition engine for both speech recognition and voice identification.

With the exception of dictation systems that use voice recognition and show text as recognized on the monitor screen, all the systems mentioned above do not provide immediate feedback to voice input.

In cable systems, several downstream data channels to receive channel and synchronization information are typically transmitted in a band of frequencies that, in the past, were reserved for re-broadcasting FM channels over cable. Currently most cable systems reserve some of the 88 to 108 MHz FM spectrum for set-top data transmission leaving the unused portion of that spectrum for

barker channels or additional video channels. The Open Cable Standard requires that the 70 to 130 MHz band be available for what's called Out-of-Band or (OOB) or Downstream transmission.

Most cable systems of today use the popular Hybrid Fiber Coax type architecture so that the downstream video signals, digital or analog, are sent to "hubs" or "nodes" via fiber optic cable. At the receiving side of the node, the optical signal from the fiber gets converted to an electrical signal containing all the analog and digital video RF carriers. This signal, in turn, is amplified and distributed via coaxial cable to all the subscriber's in the node with a typical node consisting of anywhere from 500 to 1000 subscribers. Also, the 5 to 40 MHz upstream signal from each subscriber in the node is collected, combined and then sent to the headend via the same fiber used for the downstream video carriers or a separate fiber is used.

Summary Introduction

Certain embodiments include a multi-user control system for audio visual devices incorporating a voice recognition system that is centrally located, Certain further embodiments include centrally locating the voice recognition system in or near Cable Television (CATV) headend. Certain other further embodiments include centrally locating the voice recognition system in or near a server farm. Certain other further embodiments include centrally locating the voice recognition system in or near a web-site. Certain other further embodiments include centrally locating the voice recognition system in or near a gateway.

Certain embodiments are capable of recognizing the vocal commands from a cable subscriber and then enacting upon those commands to control the delivery of entertainment and information services such as Video On Demand, Pay Per View, Channel control and the Internet. This system is unique in that the voice

command which is originated in the home of the subscriber is then sent upstream via a 5 to 40 MHz return path in the cable system to a central voice recognition and processing engine. The voice recognition and identification engine described in this disclosure is capable of processing thousands of voice commands simultaneously and therefore can offer a low latency entertainment and information experience to the Subscriber.

Functional Description (what it does)

In certain embodiments, overall media control system consists of several functional blocks with the first function being that of inputting a subscriber's voice into the system.

Figure **1** depicts a remote control unit **1000** coupled with set-top apparatus **1100** communicating via a two-stage wireline communications system containing a first wireline physical transport **1200** distributor node **1300** and a high speed physical transport **1400**, possessing various delivery points **1500** and entry points **1510-1518** to a tightly coupled server farm **2000** with one or more gateways **2100**, one or more tightly coupled server arrays **2200**, in accordance with certain embodiments.

Certain embodiments include a remote control unit **1000** fitted with microphone. Certain further embodiments include a remote control unit **1000** fitted with a special noise canceling microphone. Certain other further embodiments include a remote control unit **1000** fitted with microphone and a push-to-talk button. Certain further embodiments include a remote control unit **1000** fitted with a special noise canceling microphone and a push-to-talk button.

The purpose of the microphone in the remote is to relay the subscriber's voice commands to the central voice recognition engine. The purpose of the push-to-

talk button is to begin the process of voice recognition by informing the system that the Subscriber is about to speak and also to provide immediate address information.

In certain embodiments, voice commands from the subscriber are then preprocessed, that is the analog signals picked up from the microphone are converted to digital signals where they undergo additional processing before being transmitted to the voice recognition and identification engine located in the cable headend or other centralized location.

The preprocessing function can also take place in the remote control **1000** itself before being transmitted to the set-top box **1100** or set-top appliance **1100**, in certain embodiments.

The voice signal from the remote **1000** is a digitally modulated RF signal whose properties comply with part 15 of the FCC rules in certain embodiments. The set-top box **1100** or set-top appliance **1100** receives the voice signal from the remote **1000** and performs the preprocessing function mentioned above.

It **1100** is also used to transmit voice and subscriber address data to the centralized location or headend for voice recognition and identification. The RF signal from the remote **1000** is received by the set-top appliance **1100** and then re-modulated for upstream transmission **1200** on the 5 to 40 MHz cable return path. If a commercial set-top box **1100** is used to transmit the upstream voice data then the upstream channel allocation and transmission protocol are then controlled by the bi-directional communication system which is resident in the set-top box.

In certain alternative embodiments, a commercial set-top box **1100** is not being used to transmit the digitized voice data upstream, the set-top appliance **1100** is

then responsible for receiving the upstream channel allocation and synchronization information. The data receiver in the set-top appliance **1100** is frequency agile, that is, it can be tuned to any one of several downstream data channels to receive channel and synchronization information.

The system described specifically uses the subscriber's address information as a means by which the centrally located Agile Voice Processor can fetch a particular subscriber's parameter file. The parameter file contains voice training parameter data, voice identification parameters and user profiles for each member of the family at that address. This file can also contain parental control information and other specifics to that particular household such as language preferences or movie preferences or even internet preferences.

The Addressed Subscriber Parameter File (ASPF) is what gives the system an extremely high probability of user identification and voice recognition. Addressing is an important feature when considering secure transactions such as banking because the speech recognition and identification system has to only identify an average of 4 parameter files for any one physical address, which of course, results in a very high probability of recognizing a specific speaker's voice.

Financial level transactional security (Voice Banking) can be realized with this system and with the addition of simple voice encryption processing in the Voice Preprocessor even higher levels of security can be attained. This directly supports a method of contracting based upon an offer perceived by the user, a recognizable acceptance of the offer by an identified user.

The set-top appliance **1100** is also capable of receiving and decoding data in the downstream path. This function is required in order to organize and synchronize the transmission of upstream data. Downstream data can consist of upstream

channel allocation information and voice verification overlay information coded as text.

For embodiments where the set-top box **1100** is used for both upstream and downstream communication for the described voice command function, the function of the Set-top appliance is only to receive the RF signal from the remote control and then digitize and compress the voice signal, further preparing it for upstream transmission.

New RF protocol standards such as "Blue Tooth" allow the remote control's RF signal to transmit the voice signal directly to the set-top box where again, the preprocessing can either be done in the remote control **1000** or in firmware within the set-top box **1100**.

Set-top boxes **1100** that employ the DOCSIS type cable modems such as Open Cable set-top boxes or the so called "Heavy Set-top boxes" from Scientific Atlanta and General Instruments are capable of sending and receiving voice data using efficient data transmission protocols. The DOCSIS protocol also incorporates error detection and correction capabilities as well as other transmission enhancements such as pre-equalization for more efficient and error free transmission.

The voice signal transmitted from Subscriber's set-top box or set-top appliance **1100** is received **1510** by the 5 to 40 MHz data receiving equipment **2100** in the cable Headend.

If the digitized voice signal comes from a commercial set-top box such as a General Instruments or a Scientific Atlanta set-top, then the return path receiving equipment in the headend is specific to that type of box. Therefore, the data coming from this equipment, which will contain other upstream traffic, is parsed in

such a way that only the voice commands and address information from the subscriber are used by the AgileTV Voice Recognition Engine in the headend.

If the digitized voice signal that's being sent upstream comes from the AgileTV's set-top appliance then the upstream data receiver in the headend is a separate standalone unit designed to receive only voice command signals from the AgileTV's set-top appliance in the subscriber's home. Using the set-top appliance as the upstream transmitter allows the use of custom upstream protocols such as FM, AM, PSK or spread spectrum digital transmission. Digital transmission techniques such as QPSK or QAM can also be employed but require more costly transmission and receiver equipment.

Upon receiving the digitized and preprocessed voice signal from the subscriber's set-top box or set-top appliance, the received upstream signal will be in the form of an Ethernet data stream containing voice and address information. Since the Agile Voice Processing Unit (AVPU) is a high speed voice processing unit capable of processing the data from several nodes, the digital voice signals from each of these nodes are combined into a single high speed digital bit stream in the input multiplexer of the AVPU)

Voice Processing Engine description

Upstream signals **1510** are received at the Agile Voice Processor Unit (AVPU) RPD **2100**, in certain embodiments.

1. Voice and data signals are received from commercial return path data receivers, or:

2. Voice and data signals are received and decoded by custom return path receivers using at least one of the following protocol options: FM or AM

modulation/demodulation, FDMA, TDMA, FSK, PSK, or QPSK digital modulation/demodulation, Spread Spectrum modulation/demodulation, Telephony or cellular return or Wireless

- Application Introduction

The AVPU Engine is not an application service, in and of itself. While AgileTV may provide new end user applications, the primary function of the AVPU Engine is to provide voice control services for existing applications, such as Interactive Program Guides, and Video On Demand services.

- Application Registration

At system initialization time, applications such as VOD or Interactive Program Guides that wish to utilize voice recognition services must first register with the AVPU system. A standard program interface is utilized to enable each application to specify its complete menu hierarchy in the form of a tree structure. This tree contains "labels" for each menu, along with the text of each "button" on each menu screen. This provides the information to the AVPU engine to enable it to independently provide voice navigation services through the menu hierarchy on behalf of the application. This menu hierarchy represents the "static" portion of the application's data.

In addition to the static menu structure, it is also the responsibility of the application to inform AVPU of "dynamic" content – for example, the names of movies in a VOD system, or program names and times in an interactive program guide. Each time a user enters a menu context in which dynamic content appears, the application will inform the system of this context by passing a "handle" associated with the list of names that comprise the dynamic content. The system will combine the static menu content with the augmented dynamic

content (see Similarity Searching below), as well as application-independent keywords such as HELP, in order to form a complete "grammar". This construct is then passed to the voice recognition engine to maximize recognition accuracy.

Given that dynamic content, by definition, varies, it is the application's responsibility to inform the system whenever the content changes. In an interactive TV guide application, for example, the application may register a new set of dynamic content every one-half hour. For a VOD system, this registration would be performed whenever the database of movies offered changes.

- Application Interface

Once registration has been completed, and the system is being used, recognition of an utterance will cause a signal to be sent back to the application. This signal will inform the application to perform the requested action, and/or to update the contents of the screen as a result of the user's request. In this manner, the application can utilize the system's voice recognition services with minimal modifications to the application's code, while retaining the same graphical "look and feel" that users have become accustomed to.

- Using the System

In the subscriber's home, AgileTV supplies a voice-enabled remote control, which contains both a microphone and, in certain further embodiments, a Push-To-Talk (PTT) switch, as well as traditional universal remote control functionality. While the conventional remote control functions are transmitted via IR, voice output is transmitted as RF to a small "VoiceLink" pod located at the set-top box.

When the PTT button is pushed by the user, the remote control sends a "PTT active" command to the VoiceLink, which then informs the set-top box to place an

icon on the screen, indicating to the user that the system is "listening" to them. Next, as the user speaks into the microphone, the speech is digitized, compressed, and transmitted to the VoiceLink.

The VoiceLink encrypts the speech sample to provide security, then adds subscriber address information, a length code, and a Cyclical Redundancy Code (CRC) to enable data transmission errors to be detected.

In homes with "heavy" set-top boxes, the VoiceLink will transmit this voice information to the set-top box, which will then transmit it to the headend as a series of packets.

Otherwise, the VoiceLink will directly transmit the voice stream to the headend itself. This process continues until the VoiceLink receives a "PTT Release" from the remote, indicating end of speech. This information is also transmitted to the headend, signaling end of utterance.

- Address Decoding

Each individual consumer's interface (i.e. set-top box or set-top appliance) will have a unique address that is determined during the manufacturing process. As voice packets are transmitted upstream, this address information is pre-pended to the voice packets, enabling rapid determination of which household the voice sample is being received from. This address information is key to improving the efficiency of several different headend processing stages. The first address decode is used to assign an input buffer address to the sample. This input buffer is used to collect incoming voice packets until the final packet of a speech utterance has been received.

Certain further embodiments use on-the-fly Cyclical Redundancy Code (CRC) error checking generation. Each time a packet is read in, CRC in CPU registers is computed as each byte is read, then the partial CRC is stored at the end of the stored packet. When the next packet arrives, the partial CRC is read from where it was stored, and the new packet data is appended to the end of the previous packet, overwriting the temporary CRC. This continues until a complete voice sample has been received. By doing this, memory accesses is cut in half compared to first storing the string, then making a second pass generating the CRC.

Once a complete speech utterance has been received, the Input CPU will use the sample's source address to target the speech data to a specific voice processing CPU. This direct correspondence between the source address and a specific voice CPU is important, since it allows voice CPUs to efficiently cache user-specific parameters for the households they serve. Without this mapping, the bandwidth necessary to move household-specific data to each voice CPU would be prohibitive. In certain further embodiments, a translation table is actually used: this allows voice->CPU assignments to be changed dynamically in

the event of a hardware failure, while retaining the efficiency advantages of direct mapping.

- Grammar File is Loaded

In order for a speech sample to be processed by the voice recognition engine, it is necessary for a voice CPU to first contain in its local memory a copy of the grammar definition associated with the household's set-top box state. Speech recognition is most effective when the speech engine is aware of which words are most likely to be spoken, the order in which these words may appear, and the meaning of various sequences of words – this information is contained in a construct known as a grammar.

Before transferring the new speech sample to a voice CPU, the grammar associated with the speech sample is transferred to the target speech CPU, using a simple LRU queue. If the voice CPU contains empty space in its grammar buffer memory, then the indicated grammar is transferred directly to the empty buffer area from disk. If not, then the least-recently-used grammar buffer entry is discarded, and the new grammar information is loaded into the vacated buffer memory.

- Household Parameter File is Loaded

The next step in processing the voice sample is to ensure that the parameters associated with this household are already cached in the specific voice CPU's RAM. If these parameters are not present, then the least-recently-used parameter cache entry is evicted from the cache.

To do this, the oldest cache entry on this voice CPU is first examined to see if it has been modified. If so, the cache entry will be written to disk, and the cache

slot is then declared vacant. Next, the household speech parameters associated with the new speech sample are loaded into the vacated cache block. During the relatively long access times needed to load a new set of household parameters from disk (and optionally to write the old parameters to disk), the current voice sample will be held in the input buffer in a "waiting" state.

Only after the new household speech parameters have been loaded into the targeted voice CPU will the voice sample be moved into the work queue for the voice CPU. In this manner, the voice CPU is not held off from processing other voice requests during lengthy disk accesses. Instead, the voice CPU will continue to process other voice samples associated with households whose parameters are already in cache.

- Assignment to a Voice CPU

Once the voice parameters associated with a speech sample are finally cached in a voice CPU, the speech sample is assigned to the voice CPU by placing a descriptor for the voice sample on the target voice CPU's "work" queue. As speech samples are processed, they are removed from the front of the work queue by the voice CPU.

Eventually, the voice CPU will reach the location of the current input sample. Once this occurs, the speech sample is transferred into the voice CPU's local memory under DMA control, and the status of this voice sample is changed to "Next". This transfer occurs in parallel with the processing of the prior speech sample, ensuring that voice CPU utilization is maximized.

Once this transfer is complete, and the voice CPU completes processing of the prior sample, the status of this voice sample is changed to "Current", and the voice recognition engine begins processing this sample.

- Deadlock Elimination

Successful processing of a speech sample requires both the proper grammar and the proper household parameter information be simultaneously loaded into a voice CPU. The possibility exists of having a race condition in which a recently-loaded grammar or speech parameter file is evicted prior to its use, in the process of loading the grammar or speech parameters for the current voice sample.

To eliminate this race condition, the total number of speech samples sitting in the waiting and working queues of a voice CPU may not exceed the number of cache entries in the voice CPU.

- Speaker Identification

The first step in recognizing the current speech sample is to determine which individual person pronounced the current utterance. To do this, the "Speaker Identification" software module running on the targeted voice CPU compares the vocal characteristics of this speech sample with the characteristics of the speakers who have been previously identified in this household – these voice "templates" are an important component of the speech parameters that are cached in the CPU.

In the vast majority of utterances, the incoming speech sample will match the characteristics of a previously-identified speaker. When this occurs, the speech sample is passed on to the next phase, speech recognition.

If the speech sample is not identified with an existing speaker, then a "new user" routine is invoked, enabling a new user to be associated with this household. This routine records the new individual's speech parameters in this household's

speech parameters, so that during subsequent utterances, the new speaker will be identified by the speaker identification process.

- Speech Recognition

The inputs to the speech recognition software module are a speech sample, an individual user's speech parameters, and the grammar to be recognized. The speech engine determines the most likely utterance based on statistical analysis, and returns a text string corresponding to the utterance. As a statistical process, this matching process is probabilistic: along with the returned text string, the speech engine will also return a "percentage of match likelihood": this enables different applications to respond differently based on the calculated confidence in the recognition result.

For recognition results having a low "cost", such as a request to display listings for a particular movie, lower confidence criteria need apply. For recognition results with a high cost to the user, such as a request to purchase a movie, higher confidence thresholds may be required (furthermore, purchase verification will be requested).

When recognition accuracy is particularly low, and the voice recognition engine determined partial matches to more than one possible phrase, the engine will return the text of several possible matches. This process, known as "N-Best" enables an application, or the user, to select from several alternative recognition results.

- Voice recording

In cases where a transaction will result in a charge to the user, the system will record the voice sample representing the purchase request onto disk storage.

This voice logging is performed in order to provide verification that a purchase request was made, should a user dispute their intent to purchase. This further supports voice-based methods of contracting.

- Overlay information is generated

Due to the statistical nature of voice recognition, utterances will occasionally be mis-recognized. In fact, recognition errors are the primary hindrance to further exploitation of voice recognition systems. Recognizing this fact, this system will provide a mechanism to "close the loop" as rapidly as possible, by providing the user with immediate visual feedback regarding the recognition process. Immediately after the voice recognition engine has returned a result, visual text corresponding to the recognized utterance will be displayed on the television screen.

This is done by transmitting the recognized text string back to the set-top box. Code within the set-top box will then display the text information in a special window "on top" of the existing application display. In cases where limitations in the set-top box prevent the window from appearing on top of existing content, code within the set-top box will select a different channel containing a "standard" background field, and will display the overlay information on top of this background.

In cases where the recognition accuracy is particularly poor, and the voice engine returns several possible recognition results, this overlay display capability will be used to help refine the user's query. By displaying the text of the possible recognition results, the user can easily select from the returned list.

If none of the recognized text strings match the user's utterance, the user can always start again. However, more commonly, one of the displayed strings will

match the user's utterance. When this is the case, the user can quickly select from the displayed alternatives, enabling improved productivity in low-recognition situations.

- Similarity Searching

One key feature offered to applications is similarity searching. When the system is scanning for dynamic content (i.e. movie titles, actor names, etc), a "similarity search" feature will be used. This enables searches for names which are only partially matched, or which resemble the recognized phrase, without requiring precise specification of the exact title/name. The technique used to deliver similarity searching is to augment the dynamic content list specified by the application. Words which are similar in nature, content, or spelling to words in the content list are automatically added to the dynamic content list whenever the application informs the system of a content update. This enables the recognition grammar to be extended to support a broader array of matches without requiring bandwidth-intensive text searches each time a user's utterance has been received (links between "similar" and target words are maintained in the internal representation of the grammar, providing direct access to target names). This further enhances system performance.

Note that as used herein, similarity databases may be constructed from at least one or more of the following database architectures, relational, SQL, knowledge-based, inferential knowledge bases including Horn clause and extended Horn clause fact bases, fuzzy versions of the above, as well as neural networks and combinations of these techniques.

- Diagnostics

In order to provide robust performance, the entire AgileTV system is designed to provide fault tolerance. For individual components such as Voice CPUs and associated memory, the system will automatically recognize most component failures, and will disable faulty CPUs from being used.

In addition, most AgileTV installations are expected to require multiple AVPU boxes **2000** to meet the performance needs of the subscriber base. To ensure maximum uptime, even the total failure of an AVPU engine will not be catastrophic: the remaining AVPU engines will continue to service the incoming voice traffic, albeit at a reduced performance level. This "failover" is carried out by a negotiation protocol between the AVPU boxes.

In addition to these internal diagnostics functions, the AgileTV system also communicates continuously with AgileTV's Network Operations Center, enabling immediate identification of system errors, and rapid service dispatch to correct any problems.

System Advantages:

1. Instant visual feedback and verification.

The system is capable of overlaying text on the subscriber's TV screen immediately after a word is recognized by the system as a verification of correct or incorrect recognition.

2. Key words displayed.

The system is capable of recognizing and processing speech such that the key words of the utterance are recognized and displayed. Navigation mode only?

The system will respond to a recognized command with a visual indication of the spoken request. This visual recognition of key words can be in the form of written text or icons.

3. Latency Masking. In certain embodiments, the system is capable masking delays in upstream transmission or voice recognition. Immediately upon depressing the talk button on the remote a digital address is appended to the beginning of the voice packets to be processed. The address tells the system not only who is talking but it also provides a means for the system to begin generating masking screens or icons. Data is sent to the headend when the button on the microphone is depressed alerts the system as to who is speaking. This function allows the system to generate an icon or overlay to quickly respond to the subscriber.

4. Server centric.

The following operations are performed at a server-center located at or near the headend: Upstream transmission of voice commands. Voice command protocol. Preprocessing of voice commands-reduce data rate

5. Visual Prompting.

The use of visual prompts to stimulate fast and accurate voice navigation.

When a word is spoken and recognized the system returns the key word but also generates a list of visual prompts that can guide the subscriber to the next navigation step. By using prompts the system can appreciate optimum recognition of the prompt words thus increasing the recognition accuracy and at the same time increasing pleasure of the user's experience.

6. Adaptive Voice Recognition.

AVR is based on optimized word recognition by creating a subset of probable matches based on knowing what is in an entertainment database or the words on a Web page).

7. Secure voice transactions

This provides a method of voice based contracting across a multimedia distribution system.

8. Automatic generation of serial-multiplexed video output streams, without video boards or multiplexers!

By incorporating extremely fast, efficient arrays of micro-processors, each of which possess a frame buffer in its locally accessible memory. Each microprocessor then translates the frame buffer into an MPEG stream. Several MPEG streams may then be merged within a single microprocessor to form a multi-media stream to distribution to subscribers.

9. On the fly CRC generation

10. Efficient Similarity Searching via Augmented Grammars

11. Automatic failover via negotiated Ethernet protocols

12. Utilizing subscriber address to accelerate speaker identification, enhance security

This provides for greatly reduced latency in identifying a speaker.

Suggested Material Which Might Support Claims

1. Preprocessing of Voice Commands.

The Voice Command Preprocessor performs the function of voice filtering, digitizing, data compression," Dead Space Encoding" and address insertion. The Preprocessing function voice commands at subscriber's residence for lower upstream data rate. This processing can be located in the remote or the Set-top appliance

2. Use of Voice Packet Processor.

The Voice Packet Processor is located in the Cable Headend and is specifically designed to capture and prepare the upstream voice packets that are to be fed to the Voice Recognition Engine. The Voice Packet Processor will also perform address decoding for routing and latency reduction. It will also perform the function of decoding the "Dead Time" between words or utterances originally encoded at the preprocessor at the subscriber's side.

3. Use of a Centrally Located Natural Language Voice Recognition.

A Natural Language Voice Recognition engine computer, processor, is to be centrally located in a Cable Television Headend to process

Summary

The disclosure describes the use of voice recognition and identification in conjunction with a centrally located voice processing engine to provide a highly secure parental control system for controlling a number of video services to include but not limited to Video-On-Demand, delivered to the home via fiber or coaxial cable.

Parental Control using a centrally located voice recognition engine more specifically a voice recognition engine and associated system located in a cable television headend with the intention of serving thousands of subscribers using a single voice processing unit.

Description

This enclosure reveals a method of providing a very high level of security based on the ability of being able to precisely recognize a parents voice or a child's voice based on a limited subset of choices.

The ability to recognize a very high probability of identifying a specific speaker is based on knowing the address of the family of speakers.

More specifically, by limiting the maximum number of choices to that typically found in a subscribing family, the ability of a centralized voice recognition engine more specifically a large capacity natural voice recognition engine located in a cable headend, can readily the specific speaker allowing a parent to voice print and therefor control the type of content. Secondly, a child can learn PIN numbers but cannot fool the voice print. to identify specifically a parent or a child specific limit Specific Voice Pattern Recognition. Speaker recognition is

particularly appealing in this application, and can be applied with very high accuracy, due to the limited number of speakers per household.

1. A method that provides a superior level of parental control security by using a centrally located voice recognition engine based on knowing the address and voice prints of the speaking family.

2. A method for parents to control the specific viewing times, total time or content that a child can watch television

Headend CPU Array Architecture

Consider an array using CPU IC's **CPU**, each possessing 3 or more high speed serial protocol channels **C1-3**, a high speed interconnect bus **B1**, a low speed interconnect bus **PCI** and a local memory access bus **R**.

Figure **2** depicts a simplified block diagram using an array of CPU's as shown as **2200** in Figure **1**, in accordance with certain embodiments.

CPU1 and CPU2 are locally coupled via one of the interconnect buses. Each possesses locally accessible memory via its one **R** bus. In certain further embodiments, the R bus accessed memory is DDR SDRAM (Double Data Rate Synchronous DRAM).

The three serial protocol channels for the first CPU of each CPU pair are dedicated to communication within each row of CPU pairs. The three serial protocol channels for the second CPU of each CPU pair are dedicated to communication within each column of CPU pairs.

Assume for the moment that each CPU is operating at K billion integer operations/second, L billion floating point operations/second, the high speed serial channels support a gigabit/second bidirectional transfer rates, and that the DDR DRAMs support M gigabyte/second access rates. This system use of these integrated circuits provides 64*K billion integer operations, 64*L billion floating point operations, 128 gigabits/second interconnect throughout the array through the high speed serial channels, as well as 64*M gigabytes/second local memory access. By decoupling the tasks running on each CPU, making them independent and concurrent, this system architecture achieve the dream of MIMD computing (Multiple-Instruction Multiple Datapath), by providing essentially linear speed up with the increase in processors for up to 64 processors. Such architectures can support each CPU generating a frame in a local frame buffer and rendering it into an MPEG stream, which is then merged into a multiplexed video content serial stream suitable for a variety of multi-media applications. This apparatus is further applicable to a number of teraflop-class computing applications which today have no cost effective solution. In certain embodiments, the CPUs may each be dual instruction processors housed in a single package. In certain further embodiments, the arithmetic performance is the same as above for each processor, leading to teraflop performance. Each of these arrays will process over 1K speakers concurrently and fit into a single rack mount enclosure. Standard 72" tall racks could then hold 5 such VoiceEngines, serving up to 50,000 households concurrently. Each server contains at least dual, redundant, hot-swappable power supplies, redundant, hot-swappable disks with automatic recovery for failed disk drives. Real-time, extensive diagnostics and continuous connectivity to a 24x7 Network Operations Center ensure timely failure detection and service dispatch. In systems with more than one VoiceEngines, the server autoimaticlly handles catastrophic failures gracefully. Even if an entire VoiceEngine fails, the remaining ones pick up the workload.

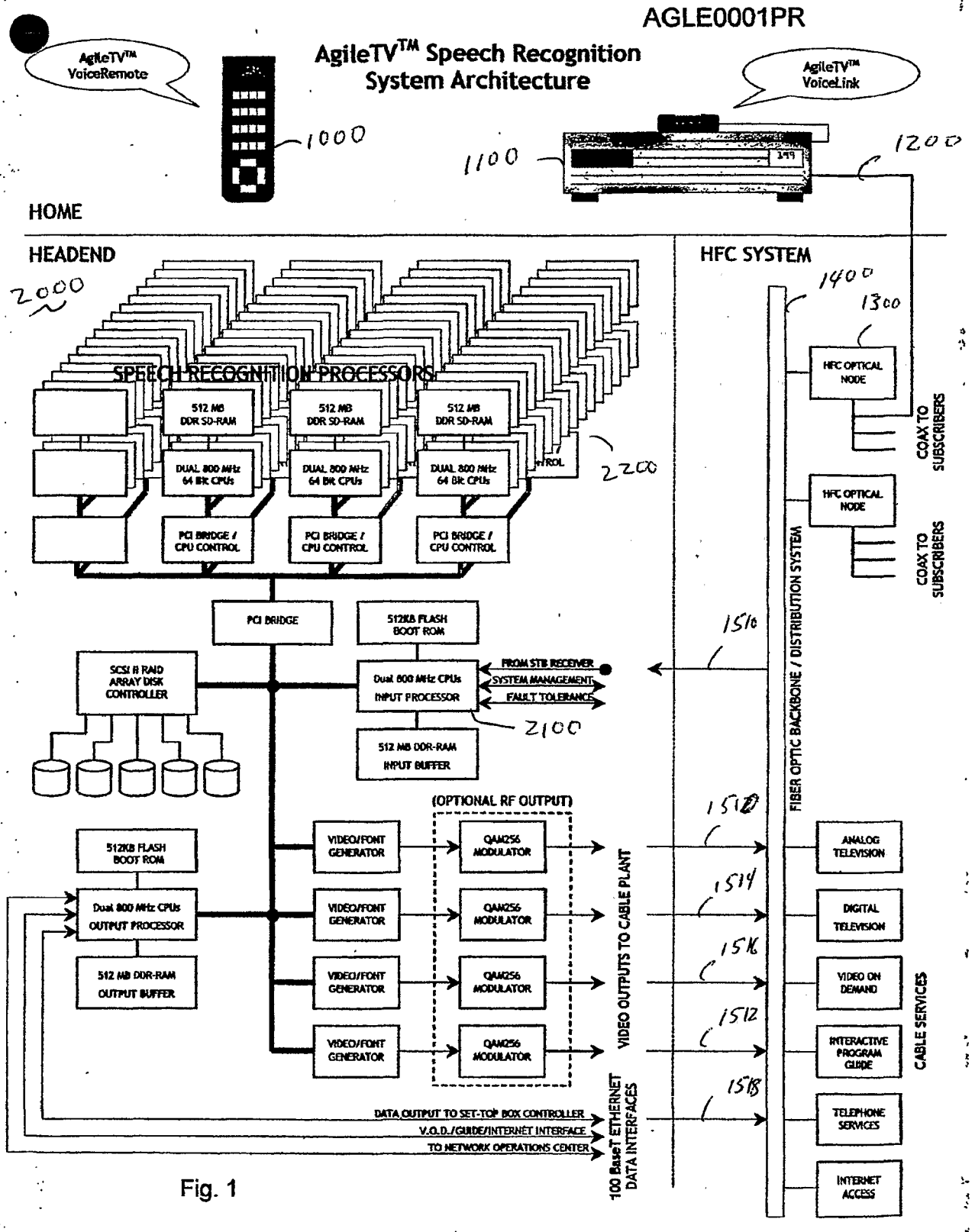
AgileTV™ Speech Recognition System Architecture
AgileTV™ VoiceRemote
AgileTV™ VoiceLink
1000
1100
1200
HOME
HEADEND
HFC SYSTEM
2000
1400
1300
SPEECH RECOGNITION PROCESSORS
512 MB DDR SD-RAM
DUAL 800 MHz 64 Bit CPUs
PCI BRIDGE / CPU CONTROL
2200
PCI BRIDGE
512KB FLASH BOOT ROM
SCSI II RAID ARRAY DISK CONTROLLER
Dual 800 MHz CPUs INPUT PROCESSOR
FROM STB RECEIVER
SYSTEM MANAGEMENT
FAULT TOLERANCE
2100
512 MB DDR-RAM INPUT BUFFER
1510
(OPTIONAL RF OUTPUT)
512KB FLASH BOOT ROM
Dual 800 MHz CPUs OUTPUT PROCESSOR
512 MB DDR-RAM OUTPUT BUFFER
VIDEO/FONT GENERATOR
QAM256 MODULATOR
VIDEO OUTPUTS TO CABLE PLANT
1512
1514
1516
1512
1518
DATA OUTPUT TO SET-TOP BOX CONTROLLER
V.O.D./GUIDE/INTERNET INTERFACE
TO NETWORK OPERATIONS CENTER
100 BaseT ETHERNET DATA INTERFACES
HFC OPTICAL NODE
COAX TO SUBSCRIBERS
FIBER OPTIC BACKBONE / DISTRIBUTION SYSTEM
ANALOG TELEVISION
DIGITAL TELEVISION
VIDEO ON DEMAND
INTERACTIVE PROGRAM GUIDE
TELEPHONE SERVICES
INTERNET ACCESS
CABLE SERVICES

Fig. 1

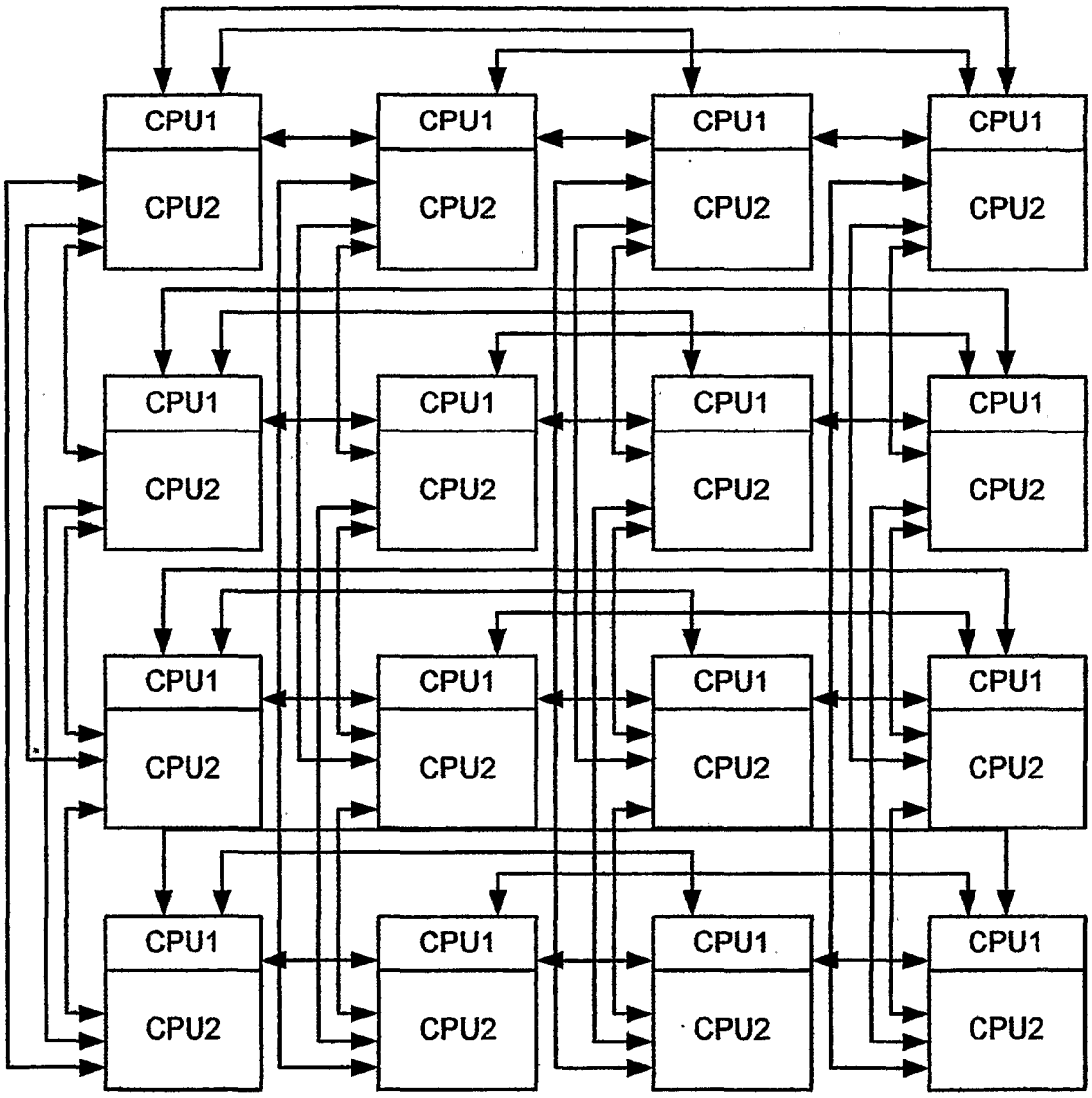

Fig. 2

AGLE0001PR

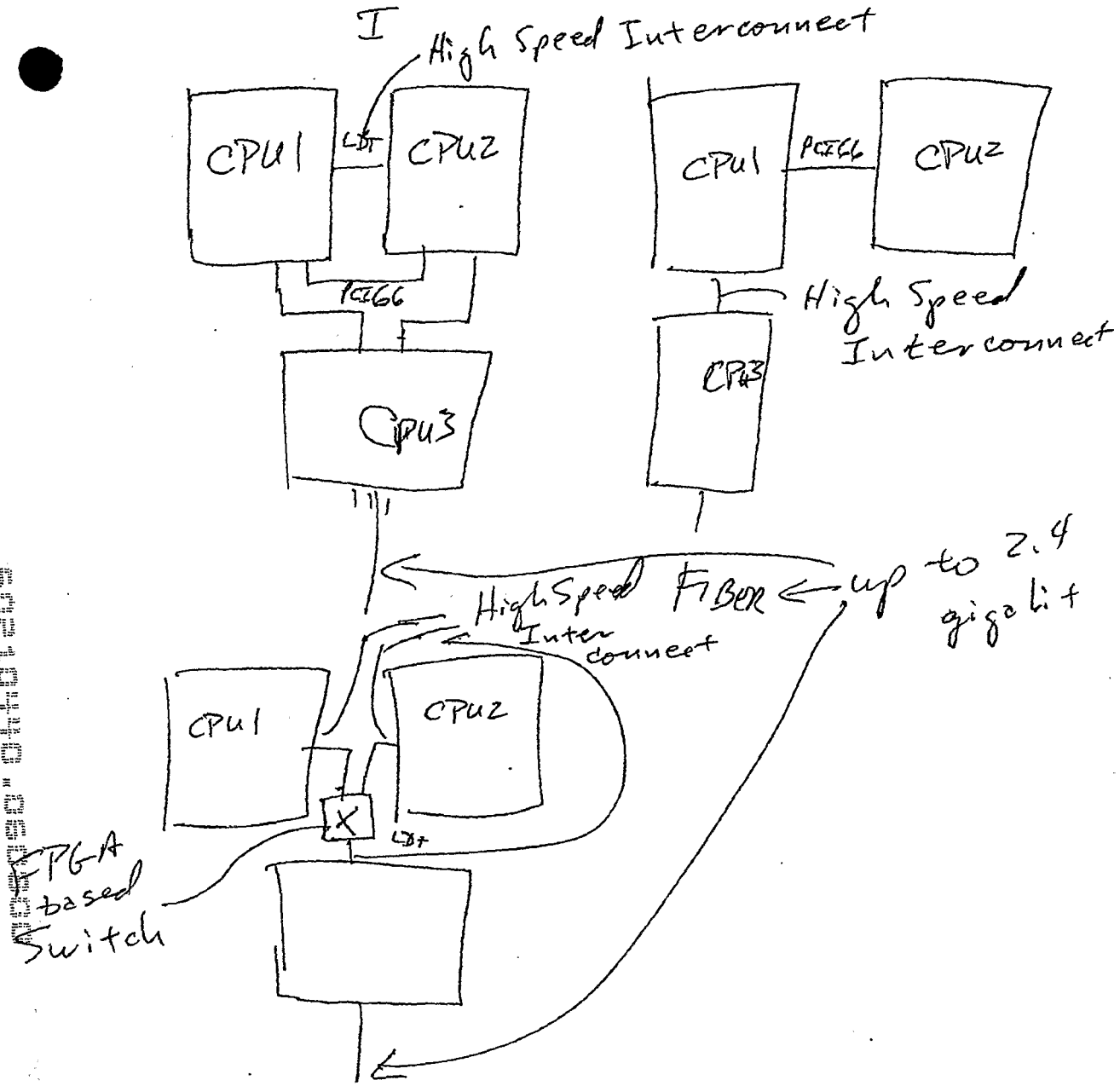

Fig 3