Iteratee IO safe, practical, declarative input processing

http://okmij.org/ftp/Streams.html

Utrecht, NL December 17, 2009

Outline

▶ Introduction

Non-solutions: Handle-based IO and Lazy IO

Pure Iteratees

General Iteratees

Lazy IO revisited

Introduction

A practical alternative to Handle and Lazy IO for input processing

Good performance

Incremental processing, interleaving, low-latency, block-based i/o from a single buffer Encouraging performance as compared to C (libsnd)

Correctness

No unsafe operations predictable resource usage, timely deallocation, preventing access to disposed resources; *Haskell98*

Elegance

Arbitrary nesting; vertical and horizontal combinations; no code bloat

http://okmij.org/ftp/Streams.html

This talk

A practical alternative to Handle and Lazy IO for input processing

- Practical talk for (server) developers
- Generalizing from practical experience (Web application server, Takusen, WAVE reader)
- ▶ Lots of code
- Use Haskell for concreteness
- Code is in Haskell98

http://okmij.org/ftp/Haskell/Iteratee/README.dr

```
PUT /file HTTP/1.1crlf
Host: example.comcr
User-agent: Xlf
content-type: text/plaincrlf
crlf
```

```
PUT /file HTTP/1.1crlf
Host: example.comcr
User-agent: Xlf
content-type: text/plaincrlf
crlf
```

```
PUT /file HTTP/1.1crlf
Host: example.comcr
User-agent: Xlf
content-type: text/plaincrlf
crlf
1Ccrlf
body line 11f body line 2crlf crlf
7crlf
body li crlf
37crlf
ne 3cr body line 4lf body line 5lf crlf
Ocrlfcrlf
```

```
PUT /file HTTP/1.1crlf
Host: example.comcr
User-agent: Xlf
content-type: text/plaincrlf
crlf
1Ccrlf
body line 11f body line 2crlf crlf
7crlf
body li crlf
37crlf
ne 3cr body line 4lf body line 5lf crlf
Ocrlfcrlf
```

```
PUT /file HTTP/1.1crlfHost:
```

```
example.comcrUser-agent: Xlf content-type: text/plaincr
```

lfcrlf1Ccrlfbody 1

ine 2crlfcrlf7

Outline

Introduction

► Non-solutions: Handle-based IO and Lazy IO

Pure Iteratees

General Iteratees

Lazy IO revisited

Non-solutions: Handle-based IO and Lazy IO

Code file: GHCBufferIO.hs

Using hGetLine, not quite correctly

```
line_read h = doread []
where
 doread acc = do
  eof <- hTsEOF h
  if eof then return (HRFail "EOF" (reverse acc))
     else do
          1 <- hGetLine h >>= return . strip_cr
          if null 1 then return (HR (reverse acc))
             else doread (1:acc)
strip_cr [] = []
strip_cr s = if last s == '\r' then init s else s
```

Using hGetLine, not quite correctly

```
line_read h = doread []
where
 doread acc = do
  eof <- hTsEOF h
  if eof then return (HRFail "EOF" (reverse acc))
     else do
          1 <- hGetLine h >>= return . strip_cr
          if null 1 then return (HR (reverse acc))
             else doread (1:acc)
strip_cr [] = []
strip_cr s = if last s == '\r' then init s else s
```

Using hGetChar

```
line_read_cr h = doread [] []
 where
 doread acc curr_line = do
  eof <- hIsEOF h
  if eof then return (HRFail "EOF" (reverse acc))
     else hGetChar h >>= check_term acc curr_line
 check_term acc curr_line '\n' = finish acc curr_line
 check_term acc curr_line '\r' = do
  eof <- hTsEOF h
  if eof then finish acc curr_line
     else do
          c <- hLookAhead h
          when (c == '\n') (hGetChar h >> return ())
          finish acc curr line
 check_term acc curr_line c = doread acc (c:curr_line)
 finish acc "" = return (HR (reverse acc))
 finish acc line = doread (reverse line:acc) ""
```

Using Lazy IO

```
line_lazy h = hGetContents h >>= return . doparse []
 where
 doparse acc str =
                         -- pure function
    case break (\c -> c == \c '\r' || c == \c '\n') str of
      ( ,"")
                     -> HRFail "EOF" (reverse acc)
       (1,'\r':'\n':rest) -> finish acc l rest
       (1,_:rest) -> finish acc 1 rest
 finish acc "" rest = HR (reverse acc)
 finish acc l rest = doparse (l:acc) rest
```

When are all resources of the Handle h freed?

Problems with Handle IO

- ▶ It is not that simple
- ► Handle IO puts the file descriptor in the non-blocking mode:
 - not always good for sockets
- Cannot do our own input multiplexing with select/epoll
- Resource leaks, closed handle errors
- Cannot do Handle IO over nested/embedded streams

Problems with Lazy IO

- ▶ It is *delusionally* simple
- Theoretical abomination:a "pure" computation with observable side-effects
- ▶ Permits no IO control
- Practically unacceptable resource management
- Practically unacceptable error reporting
- Danger of deadlocks when reading from pipes

Lazy IO in serious, server-side programming is unprofessional

Outline

Introduction

Non-solutions: Handle-based IO and Lazy IO

▶ Pure Iteratees

General Iteratees

Lazy IO revisited

Problems of the exposed traversal state

Handle exposes the (file) traversal state:

- ▶ need to pass the Handle around, and explicitly close
- danger of resource leaks or closed-Handle errors
- ▶ must check the Handle state on each access

```
fold :: (a \rightarrow b \rightarrow b) \rightarrow b \rightarrow IntMap \ a \rightarrow b

fold f z coll \equiv (f \ a_n \ ... (f \ a_2 \ (f \ a_1 \ z)))

prod = fold (*) 1 coll

\equiv (a_n \ * \ ... (a_2 \ * \ (a_1 \ * \ 1)))
```

Fold encapsulates the traversal and its resources

```
fold :: (a -> b -> b) -> b -> IntMap a -> b

fold f z coll \equiv (f a_n ...(f a_2 (f a_1 z)))

prod = fold (*) 1 coll

\equiv (a_n * ...(a_2 * (a_1 * 1)))

prodbut n = snd (fold iteratee (n,1) coll)

where iteratee a (n,s) =

if n <= 0 then (n,a*s) else (n-1,s)
```

Seed exposes the iteratee state No interface for early termination

Seed exposes the iteratee state No interface for early termination

Iteratee

```
data Stream = EOF (Maybe ErrMsg) | Chunk String
```

Iteratee

```
data Stream = EOF (Maybe ErrMsg) | Chunk String
data Iteratee a =
      IE_done a Stream
    | IE_cont (Stream -> Iteratee a) (Maybe ErrMsg)
Code file: Iteratee.hs
```

The internal 'state' of the iteratee – the seed – is fully encapsulated.

Simplest Iteratees

```
peek :: Iteratee (Maybe Char)
peek = IE_cont step Nothing
where
 step (Chunk []) = peek
 step s@(Chunk (c:_)) = IE_done (Just c) s
step stream
             = IE_done Nothing stream
head :: Iteratee Char
head = IE_cont step Nothing
where
 step (Chunk []) = head
 step (Chunk (c:t)) = IE_done c (Chunk t)
 step stream = IE_cont step (Just (setEOF stream))
```

Simplest Iteratees

```
peek :: Iteratee (Maybe Char)
peek = IE_cont step Nothing
where
 step (Chunk []) = peek
 step s@(Chunk (c:_)) = IE_done (Just c) s
step stream
             = IE_done Nothing stream
head :: Iteratee Char
head = IE_cont step Nothing
where
 step (Chunk []) = head
 step (Chunk (c:t)) = IE_done c (Chunk t)
 step stream = IE_cont step (Just (setEOF stream))
```

Complex Iteratee

```
break :: (Char -> Bool) -> Iteratee String
break cpred = IE_cont (step []) Nothing
 where
 step before (Chunk []) = IE_cont (step before) Nothing
 step before (Chunk str) =
     case Prelude.break cpred str of
       (_,[]) -> IE_cont (step (before ++ str)) Nothing
       (str,tail) -> IE_done (before ++ str) (Chunk tail)
 step before stream = IE_done before stream
```

Non-trivial state; benefiting from chunked input

Another Complex Iteratee

```
heads :: String -> Iteratee Int
heads str = loop 0 str
where
loop cnt "" = return cnt
step cnt str (Chunk "") = loop cnt str
step cnt (c:t) s@(Chunk (c':t')) =
    if c == c' then step (succ cnt) t (Chunk t')
      else IE_done cnt s
step cnt _ stream
                           = IE_done cnt stream
```

Semantics

```
"abd"...\gg heads "abc" \rightsquigarrow "d"...\gg done 2
```

Combining Iteratees

```
instance Monad Iteratee where
   return x = IE_done x (Chunk "")
   IE_done x (Chunk "") >>= f = f x
   IE_done x stream >>= f =
     case f x of
       IE_done y _ -> IE_done y stream
       IE_cont k Nothing -> k stream
                          -> i
       i
   IE_cont k e >>= f = IE_cont ((>>= f) . k) e
```

Horizontal Iteratee composition

Combining Iteratees

```
instance Monad Iteratee where
   return x = IE_done x (Chunk "")
   IE_done x (Chunk "") >>= f = f x
   IE_done x stream >>= f =
     case f x of
       IE_done y _ -> IE_done y stream
       IE_cont k Nothing -> k stream
                          -> i
       i
   IE_cont k e >>= f = IE_cont ((>>= f) . k) e
```

Horizontal Iteratee composition

Reading lines

```
type Line = String -- The line of text, no terminators
read lines :: Iteratee (Either [Line] [Line])
read lines = lines' []
where
 lines' acc = break (\c -> c == '\r' || c == '\n') >>=
      \l -> terminators >>= check acc l
 check acc _ 0 = return . Left . reverse $ acc
 check acc "" _ = return . Right . reverse $ acc
 check acc l _ = lines' (1:acc)
 terminators = heads "\r\n" >>=
   \n \rightarrow \text{if } n == 0 \text{ then heads "} n \text{ else return } n
```

Reading lines

```
lines' acc = break (\c -> c == '\r' || c == '\n') >>=
    \1 -> terminators >>= check acc 1
check acc _ 0 = return . Left . reverse $ acc
check acc "" _ = return . Right . reverse $ acc
check acc l _ = lines' (1:acc)
terminators = heads "\r\n" >>=
 n \rightarrow if n == 0 then heads "\n" else return n
doparse acc str =
                 -- for comparison
    case break (\c -> c == \c '\r' || c == \c '\n') str of
      (_,"") -> HRFail "EOF" (reverse acc)
      (1, '\r': '\n': rest) \rightarrow finish acc 1 rest
      (1,_:rest) -> finish acc 1 rest
finish acc "" rest = HR (reverse acc)
finish acc l rest = doparse (l:acc) rest
```

Reading lines

```
lines' acc = break (\c -> c == '\r' || c == '\n') >>=
    \1 -> terminators >>= check acc 1
check acc _ 0 = return . Left . reverse $ acc
check acc "" _ = return . Right . reverse $ acc
check acc l _ = lines' (1:acc)
terminators = heads "\r\n" >>=
 n \rightarrow if n == 0 then heads "\n" else return n
doparse acc str =
                 -- for comparison
    case break (\c -> c == \c '\r' || c == \c '\n') str of
      (_,"") -> HRFail "EOF" (reverse acc)
      (1, '\r': '\n': rest) \rightarrow finish acc 1 rest
      (1,_:rest) -> finish acc 1 rest
finish acc "" rest = HR (reverse acc)
finish acc l rest = doparse (l:acc) rest
```

Enumerators

```
type Enumerator a = Iteratee a -> Iteratee a
type EnumeratorM m a = Iteratee a -> m (Iteratee a)
```

Enumerators

```
type Enumerator a = Iteratee a -> Iteratee a
type EnumeratorM m a = Iteratee a -> m (Iteratee a)
(>>>):: Enumerator a -> Enumerator a -> Enumerator a
(>>>) = flip (.)
(>>.):: Monad m =>
  EnumeratorM m a -> EnumeratorM m a -> EnumeratorM m a
e1 >>. e2 = i -> e1 i >>= e2
```

Trivial Enumerators

Trivial Enumerators

```
enum_pure_1chunk :: String -> Enumerator a
enum_pure_1chunk str (IE_cont k Nothing) = k (Chunk str)
enum_pure_1chunk _ iter = iter
enum_pure_nchunk :: String -> Int -> Enumerator a
enum_pure_nchunk str@(_:_) n (IE_cont k Nothing) =
   enum_pure_nchunk s2 n (k (Chunk s1))
where (s1,s2) = splitAt n str
enum_pure_nchunk _ _ iter = iter
```

File Enumerator

```
enum fd :: Fd -> EnumeratorM IO a
enum fd fd iter =
 allocaBytes (fromIntegral buffer_size) (loop iter)
where
  buffer_size = 5 -- for tests
  loop (IE_cont k Nothing) = do_read k
  loop iter = \p -> return iter
  do_read k p = do
  n <- myfdRead fd p buffer_size</pre>
   case n of
    Left errno -> return $ k (EOF (Just "IO error"))
    Right 0 -> return $ IE_cont k Nothing
    Right n -> do
        str <- peekCAStringLen (p,fromIntegral n)</pre>
        loop (k (Chunk str)) p
```

Reading headers

```
test_driver filepath = do
  fd <- openFd filepath ReadOnly Nothing defaultFileFlags
  result <- fmap run $
            enum_fd fd read_lines_and_one_more_line
  closeFd fd
  print result
where
  read_lines_and_one_more_line = do
     lines <- read lines
     after <- break (\c -> c == ^{\prime}\r', || c == ^{\prime}\n')
     status <- is finished
     return (lines, after, status)
```

Running example

```
PUT /file HTTP/1.1crlfHost:

example.comcrUser-agent: Xlf content-type: text/plaincr

lfcrlf1Ccrlfbody 1

ine 2crlfcrlf7
```

```
type Enumeratee a = Iteratee a -> Iteratee (Iteratee a)
```

Stream nesting

- buffering,
- framing,
- character encoding,
- compression, encryption, SSL, etc.

```
type Enumeratee a = Iteratee a -> Iteratee (Iteratee a)
```

Stream nesting

- buffering,
- framing,
- character encoding,
- compression, encryption, SSL, etc.

Outer-stream elements to inner-stream elements: many-to-many

type Enumeratee a = Iteratee a -> Iteratee (Iteratee a)

```
type Enumeratee a = Iteratee a -> Iteratee (Iteratee a)
joinI :: Iteratee (Iteratee a) -> Iteratee a
joinI ii = ii >>= enum_eof
joinI ≠ monadic join
```

Simplest nesting: framing

```
take :: Int -> Enumeratee a
```

 $b_1 \cdots b_n \dots \gg$ take n $i \rightsquigarrow \dots \gg$ done i' where $b_1 \cdots b_n \gg i \rightsquigarrow \bot \gg i'$

Simplest nesting: framing

```
take :: Int -> Enumeratee a
take 0 iter@IE_cont = return iter
take n (IE_done x _) = drop n >> return (return x)
take n (IE_cont _ (Just e)) = drop n >> throwErr e
take n (IE_cont k Nothing) = IE_cont (step n k) Nothing
where
 step n k (Chunk []) = IE_cont (step n k) Nothing
 step n k chunk@(Chunk str) | length str < n =
  take (n - length str) (k chunk)
 step n k (Chunk str) = IE_done (k (Chunk s1)) (Chunk s2)
 where (s1,s2) = splitAt n str
 step n k stream = IE_done (k stream) stream
```

Chunk decoding

- ightharpoonup "O" CRLF CRLF ... \gg enum_cd i \leadsto done i
- ▶ n_{hex} CRLF $b_1\cdots b_n$ CRLF \ldots \ggg enum_cd i \leadsto \ldots \ggg enum_cd i' where $b_1\cdots b_n$ \ggg i \leadsto $_$ \ggg i'

Chunk decoding

```
enum chunk decoded :: Enumeratee a
enum_chunk_decoded iter = read size
where
read_size = break (== '\r') >>=
             checkCRLF iter . check size
 checkCRLF iter m = do
  n \leftarrow heads "\r\n"
   if n == 2 then m else frame err "..." iter
 check size "0" = checkCRLF iter (return iter)
 check size str@( : ) =
     maybe (frame_err "Chunk size" iter) read_chunk $
     read hex 0 str
 check_size _ = frame_err "Error reading chink size" iter
 read_chunk size = take size iter >>= \r ->
   checkCRLF r $ enum_chunk_decoded r
```

Complete test

```
test_driver filepath = do
  fd <- openFd filepath ReadOnly Nothing defaultFileFlags
  result <- fmap run (enum_fd fd read_headers_body)</pre>
  closeFd fd
  print result
where
  read_headers_body = do
     headers <- read_lines
     body <- joinI (enum_chunk_decoded read_lines)</pre>
     status <- is_finished
     return (headers, body, status)
```

Running example

```
PUT /file HTTP/1.1crlfHost:

example.comcrUser-agent: Xlf content-type: text/plaincr

lfcrlf1Ccrlfbody 1

ine 2crlfcrlf7
```

Outline

Introduction

Non-solutions: Handle-based IO and Lazy IO

Pure Iteratees

▶ General Iteratees

Lazy IO revisited

General Streams and Iteratees

```
data Stream el = EOF (Maybe ErrMsg) | Chunk [el]
data IterV el m a =
     IE_done a (Stream el)
   | IE_cont (Stream el -> Iteratee el m a) (Maybe ErrMsg)
newtype Iteratee el m a =
      Iteratee{runIter:: m (IterV el m a)}
instance Monad m => Monad (Iteratee el m)
instance MonadTrans (Iteratee el)
```

Code file: IterateeM.hs

Sample General Iteratees

```
head :: Monad m => Iteratee el m el
break :: Monad m => (el -> Bool) -> Iteratee el m [el]
dropWhile :: Monad m =>
  (el -> Bool) -> Iteratee el m ()
drop :: Monad m => Int -> Iteratee el m ()
line :: Monad m => Iteratee Char m (Either Line Line)
stream2list :: Monad m => Iteratee el m [el]
print_lines :: Iteratee Line IO ()
```

```
type Enumerator el m a = IterV el m a -> Iteratee el m a \cong IterV el m a -> m (IterV el m a)
```

```
type Enumerator el m a = IterV el m a -> Iteratee el m a \cong IterV el m a -> m (IterV el m a)
```

```
Why not the following type?
```

```
type Enumerator el m a = 

Iteratee el m a -> Iteratee el m a 

\cong m (IterV el m a) -> m (IterV el m a)
```

Troublesome code:

```
do let iter = enum_file file1 iter_count
   some_action
   run (enum_file file2 iter)
```

```
type Enumerator el m a = IterV el m a -> Iteratee el m a \cong IterV el m a -> m (IterV el m a)
```

```
Why not the following type?
```

```
type Enumerator el m a = 

Iteratee el m a -> Iteratee el m a 

\cong m (IterV el m a) -> m (IterV el m a)
```

Troublesome code:

```
do let iter = enum_file file1 iter_count
   some_action
   run (enum_file file2 iter)
```

```
type Enumerator el m a = IterV el m a -> Iteratee el m a \cong IterV el m a -> m (IterV el m a)
```

```
infixr 0 $$
f $$ x = x >>== f

(>>>):: Monad m =>
    Enumerator el m a -> Enumerator el m a ->
    Enumerator el m a
-- (>>>) = flip (.)
e1 >>> e2 = \i -> e2 $$ (e1 i)
```

Sample General Enumerators

```
enum_eof :: Monad m => Enumerator el m a
```

 $enum_fd :: Fd \rightarrow Enumerator Char IO a$

Sample General Enumeratees

```
type Enumeratee elo eli m a =
    IterV eli m a -> Iteratee elo m (IterV eli m a)

take :: Monad m => Int -> Enumeratee el el m a
enum_chunk_decoded :: Monad m => Enumeratee Char Char m a
```

Sample General Enumeratees

```
type Enumeratee elo eli m a =
    IterV eli m a -> Iteratee elo m (IterV eli m a)

take :: Monad m => Int -> Enumeratee el el m a
enum_chunk_decoded :: Monad m => Enumeratee Char Char m a
```

```
joinI :: Monad m =>
    Iteratee elo m (IterV eli m a) -> Iteratee elo m a
joinI outer = outer >>= lift . run
```

Sample General Enumeratees

```
type Enumeratee elo eli m a =
    IterV eli m a -> Iteratee elo m (IterV eli m a)
take :: Monad m => Int -> Enumeratee el el m a
enum chunk decoded :: Monad m => Enumeratee Char Char m a
joinI :: Monad m =>
   Iteratee elo m (IterV eli m a) -> Iteratee elo m a
joinI outer = outer >>= lift . run
infix1 1 >>==
(>>==):: Monad m => Iteratee el m a ->
         (IterV el m a -> Iteratee el' m b) ->
         Iteratee el' m b
m >>== f = Iteratee (runIter m >>= runIter . f)
```

More interesting Enumeratees

```
map_stream :: Monad m =>
    (elo -> eli) -> Enumeratee elo eli m a
enum_lines :: Monad m => Enumeratee Char Line m a
sequence_stream :: Monad m =>
    Iteratee elo m eli -> Enumeratee elo eli m a
```

True IO interleaving

```
line_printer = enum_lines $$ print_lines
print_headers_print_body = do
     lift $ putStrLn "Lines of the headers follow"
     line_printer
     lift $ putStrLn "Lines of the body follow"
     joinI $ enum_chunk_decoded $$ line_printer
test_driver_full iter fpath = do
  fd <- openFd fpath ReadOnly Nothing defaultFileFlags
  run $ enum_fd fd $$ iter
  closeFd fd; putStrLn "Finished reading"
test_driver_mux iter fpath1 fpath2 = do ...
```

Outline

Introduction

Non-solutions: Handle-based IO and Lazy IO

Pure Iteratees

General Iteratees

▶ Lazy IO revisited

Lazy IO vs. Iteratee IO

```
driver1 (i:j:rest) =
    print (max_cycle_len i j) >> driver1 rest
driver1 _ = return ()
main1 = getContents >>= driver1 . map read . words
```

Code file: GetContentsLess.hs

Lazy IO vs. Iteratee IO

```
driver1 (i:j:rest) =
   print (max_cycle_len i j) >> driver1 rest
driver1 = return ()
main1 = getContents >>= driver1 . map read . words
driver2 = do
          i <- head; j <- head
          lift (print (max_cycle_len i j)) >> driver2
main2 = run $
  enum_file "/dev/tty" $$
  (enum_words $$ map_stream read $$ driver2)
```

Code file: GetContentsLess.hs

Binary and random IO

RandomIO.hs

Reading 16- or 32-bit signed and unsigned integers in big- or little-endian formats; Seeking within a file

Tiff.hs

An extensive example of:

- random and binary IO;
- on-demand incremental processing with iteratees.

Conclusions

Iteratee IO: safe and practical alternative to Lazy and Handle IO

- Compositionality
 - Iteratees compose horizontally as monads
 - Iteratees compose vertically: nesting, embedded stream processors
 - Enumerators are iteratee transformers, compose as functions
- Good resource management
- Good error handling
- Inherent incremental processing
- Safe IO interleaving
- Based on left fold, for any FP language

Good performance, Correctness, Elegance