Audio Rendering Tag

IETF CLUE Interim Meeting June 2012

Brian Baldino
bbaldino@cisco.com
Rob Hansen
rohanse2@cisco.com
Allyn Romanow
allyn@cisco.com

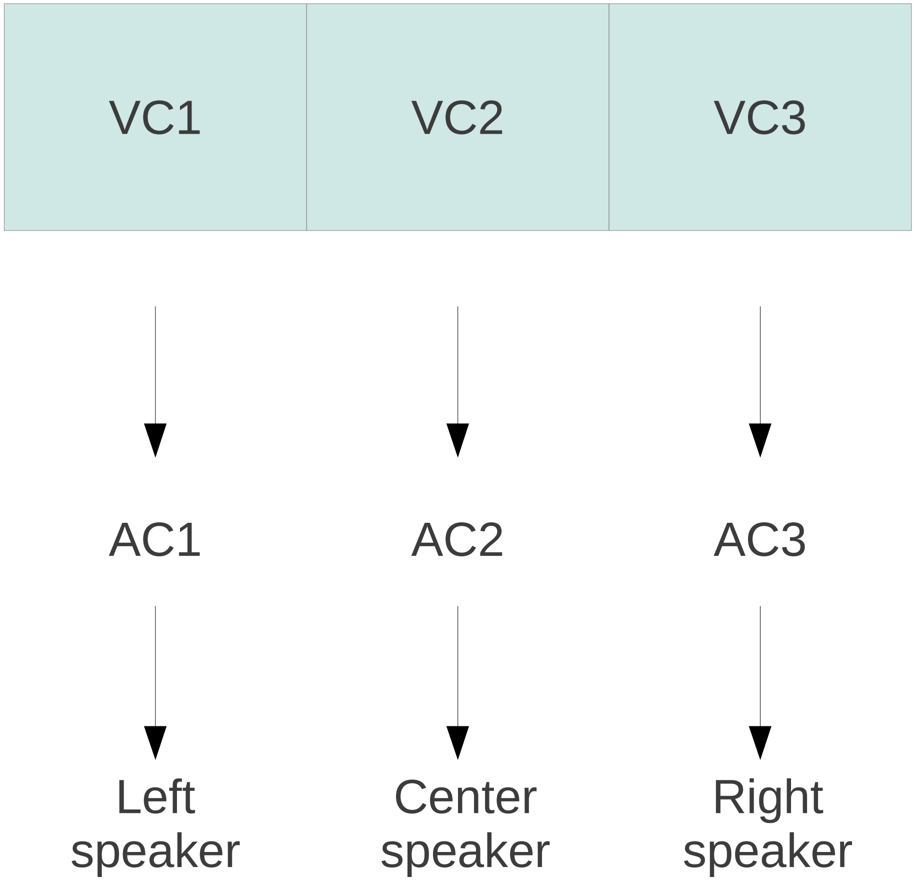
Motivation

- Stems from the need to support “directional audio”
  - Playing out to potentially multiple loud speakers at the consumer; loud speaker positions are only known by the consumer
    - Gives better, more immersive, Telepresence experience
  - Number of loud speakers not necessarily related to the number of decoded streams
  - Not the same as or related to lip sync
    - Lip sync taken care of by RTCP CNAME

Implications of switching

- If “ordered speaker” switching captures in use, mapping between received VC1 .. VCn and AC1 .. ACm could be very fluid
  - Consumer would need to dynamically redirect received AC1 .. ACm to different loudspeakers as active participants change
  - Sometimes “top M” audio streams will include placed (visible) participants, sometimes not
    - Want to avoid the need for consumer → provider CLUE message to be sent with information on loud speaker positions due to the frequency of changes

3 screen endpoint example



VC1 VC2 VC3

AC1 AC2 AC3

Left speaker

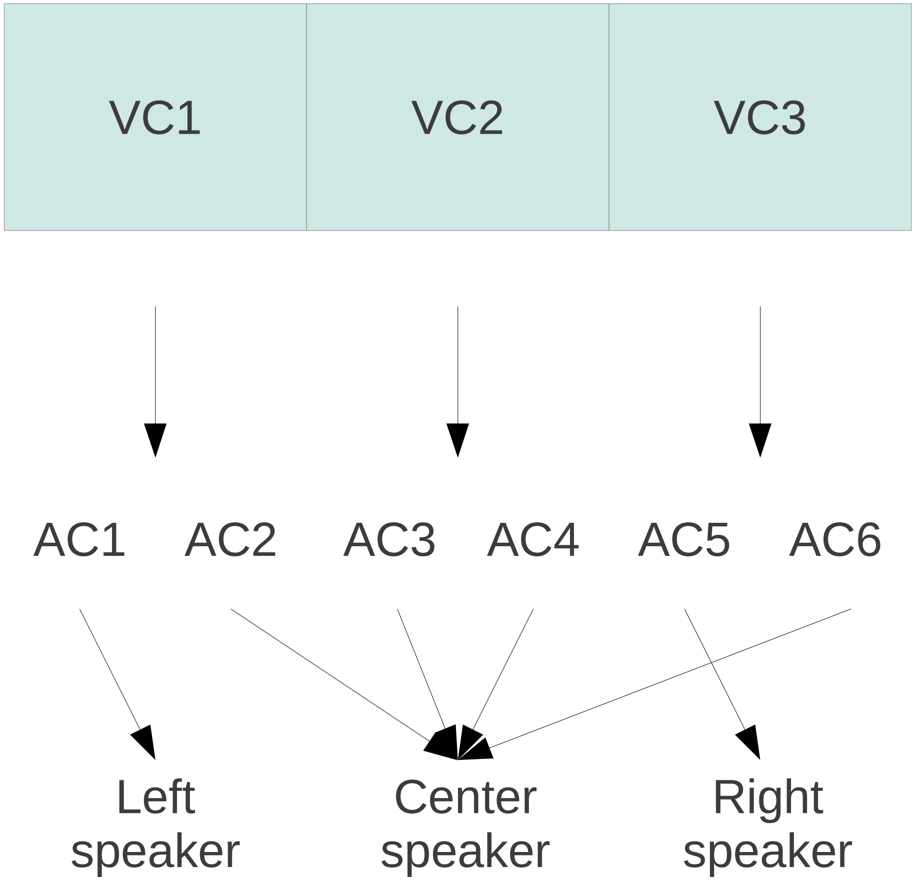
Center speaker

Right speaker

Video captures
(one per screen)

Audio captures
(one per speaker)

3 screen endpoint example



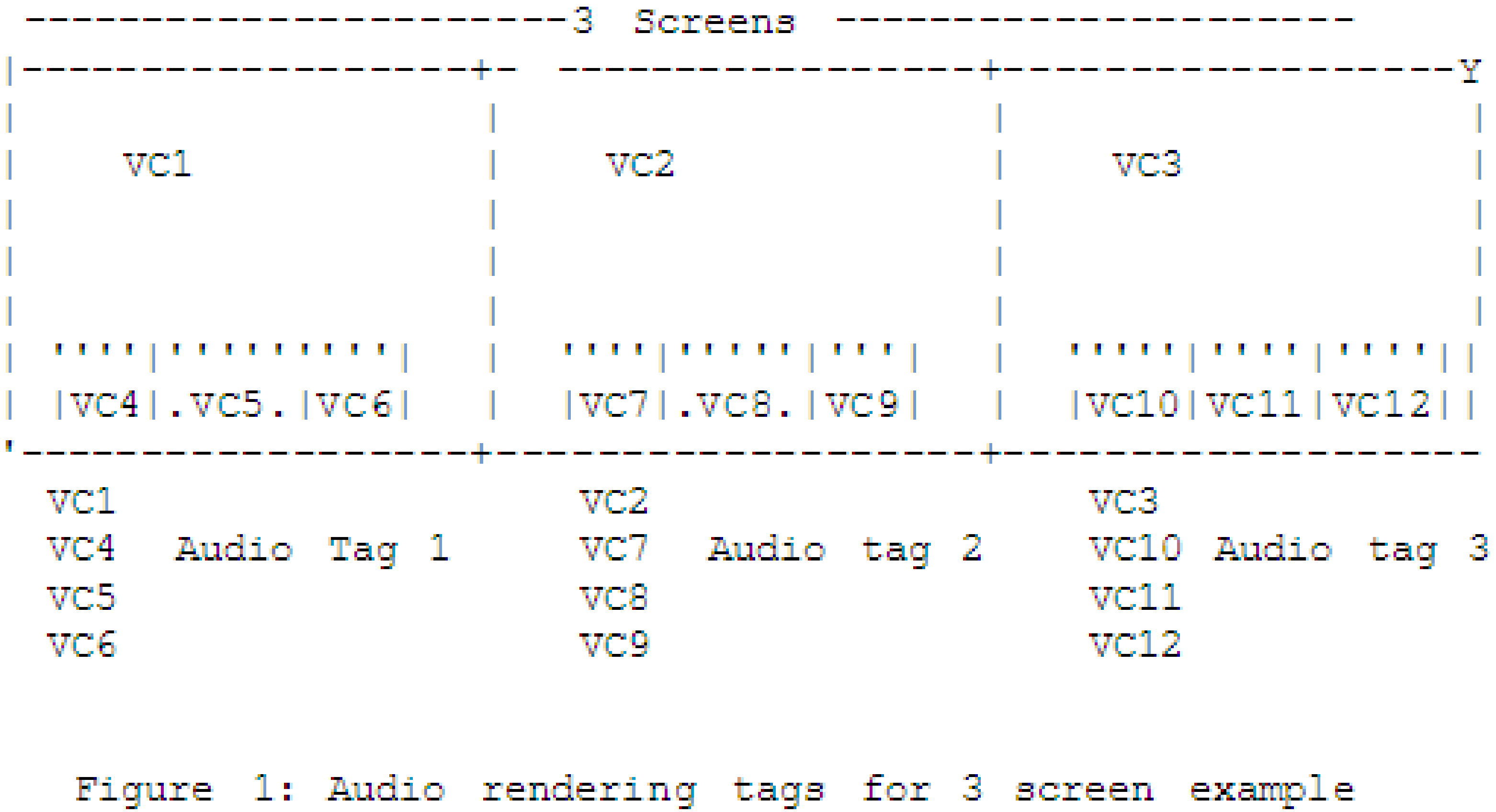
Video captures; one per screen. Number of received video captures may be different from number of audio captures.

Audio captures; AC1 is associated with VC1, AC5 with VC3, and other audio captures have no directional association.

Audio tagging scheme

- Idea is for consumer to supply an “audio tag” value for each video capture it chooses to receive
  - Provider tags audio captures corresponding to those video captures with specified audio tag
    - Audio tag values implemented with an RTP header extension
  - Consumer uses received audio tag values to direct decoded media streams to appropriate loud speaker
    - Audio captures not corresponding to a selected video capture will not have a tag – consumer will fall back to “default” behaviour; e.g. a central speaker

3 screen audio tagging example



```
 ----------------------3  Screens  ---------------------
|------------------+-  ----------------+------------------Y
|                  |                   |                  |
|    VC1           |    VC2            |    VC3           |
|                  |                   |                  |
|                  |                   |                  |
|                  |                   |                  |
| ''''|''''''''''| |   ''''|''''''|''''| |   '''''|'''''|'''''||
| |VC4|.VC5.|VC6|  |   |VC7|.VC8.|VC9|  |   |VC10|VC11|VC12||
'------------------+-------------------+------------------
  VC1                VC2                  VC3
  VC4   Audio Tag 1  VC7   Audio tag 2    VC10  Audio tag 3
  VC5                VC8                  VC11
  VC6                VC9                  VC12


   Figure 1: Audio rendering tags for 3 screen example
```

Consumer could vary its behavior, for instance, choosing whether VC4–VC12 were significant enough to have spatial audio component

Other cases to consider

VC1

VC2

VC3

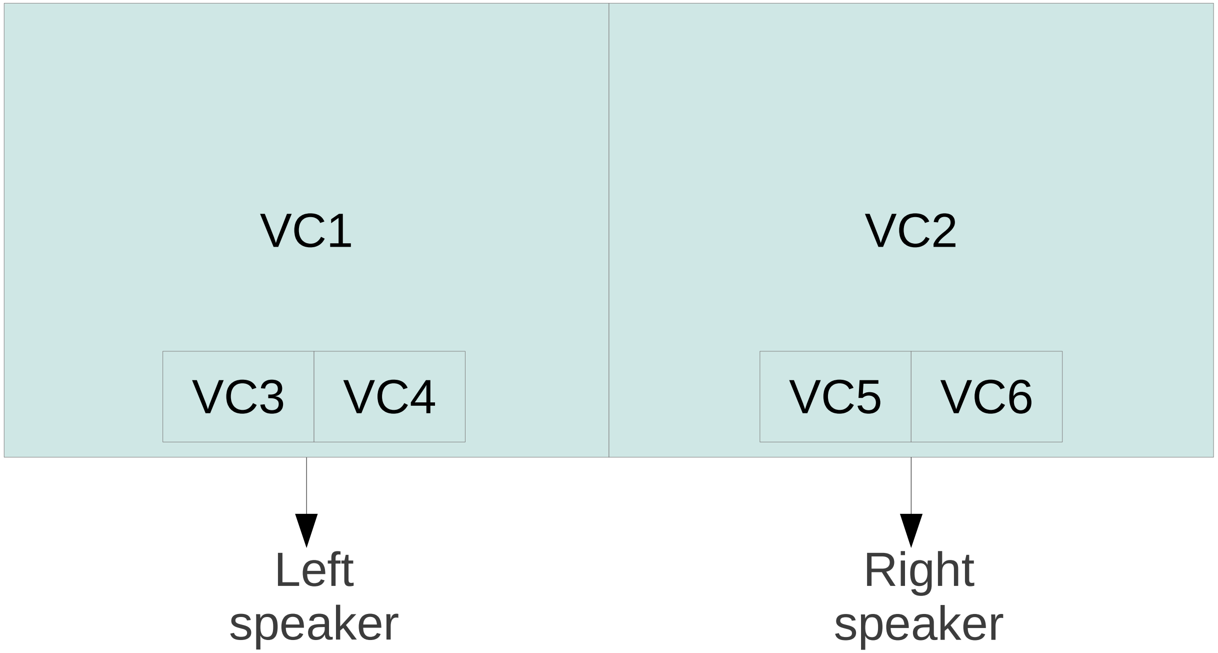
“One to many”
What if VC1 and VC2 are 2 camera system contributing a single mono audio capture?

- Corresponding AC<n> could have multiple tags
- Provider might only add audio tag if unambiguous
- Audio tag values could be defined to be meaningful if summed

“Many to one”
What if VC1 contributes separate L / C / R audio captures?

- all 3 audio captures received with VC1 tag value
- Source spatial audio “compressed” to single speaker output

2 screen example



Want to associate VC1, VC3 and VC4 with left speaker, and VC2, VC5 and VC6 with right speaker